ECOLOGICAL GENETICS OF ARABIDOPSIS THALIANA FROM RESERVOIR POPULATIONS IN LOW-DISTURBANCE HABITATS

Neil Pearson

Doctorate of Philosophy

University of Warwick, School of Life Sciences

September 2013

THE UNIVERSITY OF
WARWICK

Table of Contents

LIST OF FIGURES AND TABLES

FIGURES

TABLES

ACKNOWLEGEMENTS

This project would not have been possible without the patient support, excellent advice and great knowledge of many people within and outside of the University.

I would like to thank Prof. Eric Holub for supervising this project. He has been a greatly reassuring and wise influence at every stage, from the very beginning when he bought me some lunch when I came for an interview, through guiding my research towards the most effective and interesting implementation possible, all the way to editing out all my crimes against grammar committed in late-night write-up sessions.

I would also like to thank Dr. Robin Allaby, who co-supervised the project. From his combination of a remarkable insight and an astonishing breadth of knowledge have arisen many of the cleverest and most fascinating components of this project. I invariably left his company itching to put some new and exciting idea into practice.

Thanks are also due to Dr Volkan Cevik and Andrew Mead, who lent their adeptness with plant/pathogen interactions and statistics respectively, and who made both of these forays into what was, at the time, relatively unfamiliar territory an absolute joy.

My appreciation goes out to Matthew Horton for sharing his lab's data on meiotic crossover rates, and to Alexander Platt for the extensive and helpful advice he gave me when we met at SMBE in Chicago.

I am grateful to all of the other PhD students and postdocs who have become friends in the time I have spent here, and especially to my family and my wonderful wife, for keeping me sane and happy even through the most stressful points.

Finally, I would like to thank the BBSRC, for funding this whole business.

GLOSSARY OF TERMS

Adaptation: The process by which populations change to become better able to survive in the range of habitats in which they exist.

Allele: An alternative form of a gene, in which some nucleotides in the sequence are different from other copies of the gene extant in a population. Different alleles may cause different phenotypes.

ANCOVA: Analysis of covariance; a statistical test to determine the extent to which changes in the observations of two variables are linked to each other.

Assembly: The process of joining a large number of overlapping short DNA sequences together into (ideally) a single continuous set of sequence representing the complete genome of an organism.

Balancing selection: Selection that acts to maintain multiple alleles in the gene pool of a population. This situation most commonly arises when either heterozygotes or rare alleles are favoured by natural selection.

Bottleneck: An event in which a species or population is temporarily reduced to a small number of individuals. Rare alleles are likely to go extinct, and genetic drift becomes a more significant factor in the rate of change in allele frequencies between generations.

Chiasma (crossover): Locus at which recombination between homologous chromosomes occurs in meiosis.

Demographics: Characteristics of the typical distribution of individuals across a population – the term may relate to typical patterns of movement or the presence of organisms in a given set of circumstances.

Dispersal: The movement of seeds or pollen from one location to another.

Effector: After initial infection, hosts may activate defence pathways upon the detection of pathogen-associated molecular motifs (see **PAMP-triggered immunity**). Effector proteins produced by the pathogen inhibit the activation of the host's defence pathways, triggering a state of **effector-triggered susceptibility (ETS)**. Effectors may also have evolved to inhibit the action of R proteins.

Effector-triggered immunity (ETI): Pathogens may release effectors to inhibit host defence pathways. This inhibition is detected by **R proteins**, which induce strong defences – including the **hypersensitive response** – that result in a high degree of resistance to further infection.

Fitness: The degree to which an individual is able to survive and reproduce in a specified set of circumstances.

Founder effect: A special case of a bottleneck, in which a small number of individuals colonising an isolated and previously uninhabited habitat experience the extinction of rare alleles and increased influence of random sampling due to the small size of the local population.

Gene: A region of an organism's genome responsible for encoding a single protein.

Gene flow: The transfer of alleles from one area or sub-population to another, either by – in the case of plant species – dispersal of seeds, or by dispersal of gametes.

Gene Ontology (GO): A controlled language framework in which the characteristics of genes and proteins are described using terms with specific, consistent, predetermined meanings.

Genetic drift: Variation in the frequencies of alleles from generation to generation caused by random sampling. Drift is a more significant factor in the rate of change of allele frequencies in smaller populations than in larger ones.

Genome-Wide Association Study (GWAS): A genetic study that attempts to find alleles responsible for variation in a given phenotype by testing for statistical associations between that phenotype and a genome-wide set of molecular markers (typically SNP alleles).

Genotype: The set of alleles an individual possesses.

Habitat: A location and/or set of biotic and abiotic conditions that a population may inhabit.

Habitat type: Within this project, "habitat type" is used to refer to a specified set of environmental conditions, which may be found reasonably consistently across multiple different locations.

Haplotype: A set of alleles that consistently appears together as a group in the genotypes of a population. Recombination causes haplotypes to break apart over time; that is, the consistency of the set of alleles becomes reduced over successive generations.

Haplotype block: A haplotype spread across multiple individuals. The haplotype may be broken apart to a different extent in each individual, but the set of haplotypes as a whole may still be referred to as a haplotype block.

Hardy-Weinberg equilibrium: A population model describing the circumstances under which the frequency of genotypes does not change between generations. When a change in genotype frequencies is observed, it follows that the population differs from the model in at least one of those circumstances.

Hit (as in, hit data): Clustering analysis was applied individually to genotype subsets (referred to as windows). A single genotype cluster at a single window is referred to as a 'hit'. Hits in adjacent windows were then joined together, if possible, thus identifying haplotypes.

Hypersensitive response: The programmed, rapid death of cells in the vicinity of pathogens, triggered by the activation of **R proteins**. Lysed cells may also release compounds and enzymes that damage pathogens, or inhibit their reproduction. This localised cell death prevents the proliferation of biotrophic pathogen species by limiting their food supply, but may increase the host's susceptibility to necrotrophic pathogens.

Isolation by distance (IBD): A type of population structure in which the likelihood of individuals sharing a genotype is inversely proportional to the distance separating them.

Leucine-rich repeat (LRR): A protein structure motif found in proteins with a diverse range of functions. Proteins with this motif frequently function in the identification of other proteins, including proteins associated with the presence of pathogens (i.e., acting as **Pattern recognition receptors**). R genes may also possess LRR domains. Genes with LRR motifs that also possess nucleotide-binding capability are abbreviated as NB-LRR genes.

Linkage disequilibrium (LD): An association between two alleles at different loci at a rate significantly differing from chance. An allele at once locus may be found exclusively in the presence of an allele at a second locus, for example. This indicates that one or more of natural selection, population structure or random sampling effects are occurring.

Local adaptation: The process by which populations change to become better able to survive in a specific habitat, potentially at the expense of fitness to other habitats.

Locus: A specific, consistent location along the length of a chromosome or genome. The precise position of a gene along the length of a chromosome is described as its locus.

Meiosis: The cell division event responsible for reducing the diploid number of chromosomes to the haploid number, thereby producing gametes for the purpose of sexual reproduction. Recombination occurs as a consequence of the halving in chromosome number.

Metapopulation: The overall term for a group of populations separated by spatial distance or other barriers to gene flow, between which some degree of interaction occurs.

Outcrossing: Sexual reproduction with a non-self individual of the same species.

PAMP-triggered immunity (PTI): Immunity triggered by the detection of molecular motifs indicating the presence of pathogens (**PAMPs**). The Zigzag model describes how initial infection of a host by a pathogen may be recognised by pattern recognition receptor proteins, triggering a network of defence signals resulting in a moderate degree of resistance to further infection.

Pathogen-associated molecular pattern (PAMP): A conserved molecular motif associated with the presence of a pathogen. Triggers the innate immune response (**PAMP-triggered immunity**) when recognised by **Pattern recognition receptor** proteins in plants (see **Zigzag model**). **Microbe-associated molecular pattern (MAMP)** is often used almost interchangeably in the literature.

Pattern recognition receptor (PRR): Proteins that recognise molecular motifs (**PAMPs**) indicating the presence of pathogens. Recognised patterns include bacterial flagellin and peptidoglycan, fungal chitin, and viral nucleic acids (e.g. double-stranded RNAs). PRR proteins trigger defence signaling pathways of the innate immune system.

Phenotype: The physical characteristics of an organism upon which natural selection may act.

Population: A group of organisms; commonly defined as those inhabiting a specified area, belonging to a species, or both.

Population genetics: The branch of genetics that studies how the frequencies of alleles across populations arise and change over time, and proposes frameworks for understanding the causes of those changes.

Population structure: Consistent, non-random choice of mates for sexual reproduction within a population. In the case of a metapopulation of isolated sub-populations, individuals are more likely to mate with other individuals within their local group. In the case of isolation by distance, individuals are more likely to mate with others located physically near to them.

Principal Coordinate Analysis (PCA) (also known as Multidimensional Scaling): A statistical analysis technique designed to demonstrate the similarity of individual cases to each other. Variation between cases is reframed in terms of difference between cases, rather than in dimensions measured by experiment.

R gene/R protein: A gene product involved in **effector-triggered immunity**. Proteins produced by R genes (R proteins) detect the influence of pathogen effector proteins on the host's defence pathways (the **guard hypothesis**), and respond by directly inducing strong defence responses, including the **hypersensitive response**.

Recombinant Inbred Line (RIL): A lineage created from a single individual produced in a genetic crossing experiment after 2 or more generations of backcrossing (i.e., a lineage descended from an F2 or greater individual).

Recombination: The process of homologous chromosomes becoming physically joined at multiple loci during meiosis. This ensures that the diploid number of chromosomes is reduced to the haploid number correctly. A consequence is that sets of alleles are interchanged between homologous chromosomes.

Selection (or natural selection): The process through which individuals with a greater degree of fitness to existing habitats and conditions produce

Selective sweep: Selection that acts to cause a single allele to increase in frequency in the population until all other alleles become extinct, at which point the allele under selection is said to have reached fixation.

Selfing (or self-fertilisation): Sexual reproduction in which both gametes are supplied by the same individual.

Single nucleotide polymorphism (SNP): An allele in which a single base in a DNA sequence is altered.

Window: In this project, a dense SNP dataset was divided into a set of thousands of smaller sets of SNP alleles at contiguous loci, each covering a small section of the whole genome. Each set of SNPs is termed a window.

Zigzag model: This model represents plant-pathogen interactions as 4 stages. Firstly, the host detects the presence of pathogens via **PAMP-triggered immunity**. Secondly, pathogens may suppress PAMP-triggered immunity via **effectors**, inducing susceptibility in the host. Thirdly, R proteins detect inhibition caused by effectors and induce **effector-triggered immunity** responses, including the **hypersensitive response**. Fourthly, a second effector may suppress effector-triggered immunity, which in turn may be recognised by a second R protein.

SUMMARY

The Arabidopsis HapMap project, and follow-on work carried out by the Bergelson and Nordborg groups, established in broad outline the demographic history and population structure of wild *Arabidopsis thaliana*. Genome-wide association studies are likewise making considerable advances in identifying genes associated with ecologically significant traits, and thus in identifying candidate genes likely to be under the action of natural selection.

The aim of this project has been to further expand and combine these lines of investigation, by using genomic data to test ecological hypotheses and to grant more complete insight into the range of selection pressures acting upon wild populations. A method to measure and elucidate the genetic similarity of genomic regions between sampled accessions was therefore developed to facilitate this. 250K SNP data from RegMap accessions was then examined for evidence of patterns of migration and gene flow across Europe. Those observations formed the basis of a simple model of the history of the UK population relative to that of Europe. Comparisons of observed genotypes against expectations derived from the model allowed the identification of genomic regions under the influence of selection. Loci corresponding to signatures of selection indicated positive selection acting upon phenotypes of disease resistance, flowering time, and seed size.

CHAPTER 1: INTRODUCTION

1.1 THE PAST AND THE FUTURE

We live in a genomic age.

Scientists today hold in their hands a wealth of data unimaginable even a few decades ago. With a moment's search, a scientist can now retrieve practically any paper published in the last half century, or access the full results of any experiment on record, or follow a line of thought or the growth of an idea back through time to its originator or forward to the very latest relevant developments introduced by other scientists around the world.

Nowhere has the effect of this unprecedented flood of information had more of an impact than the life sciences. Just as the Industrial Revolution manifested humanity's growing mastery of the sciences of physics and chemistry, so our growing ability to penetrate – and turn to our advantage – the mysteries presented to us by living organisms marks a transition towards our mastery of the science of biology. As our knowledge grows, we find answers to long-standing questions, and also how best to ask new ones.

Some of these questions are fundamental to the understanding of life. How does variation within a gene affect the physical characteristics of an organism that are of ecological importance? Which genes – and which alleles of those genes – does natural selection act upon, or not act upon, and why?

Other questions apply to more specific circumstances, though may also serve as examples from which more widely applicable principles may be derived. Does the evolution of genes involved in the defence against disease proceed through novel alleles spreading rapidly through a population, or through the preservation of genetic diversity? To what extent do phenotypic plasticity and genetic responses to eternal stimuli contribute to overall fitness? As the climate

changes across the globe, how might we expect populations of organisms in the wild to adapt? (See (Mitchell-Olds et al. 2007)).

Yet on the other hand, at no time has there been a more pressing need for a sound understanding of key ecological processes. Every ecosystem on Earth is now influenced by human action (Vitousek 1997). A worryingly large number of those ecosystems are threatened with unprecedented damage or even total collapse in the foreseeable future Cos(Jackson et al. 2001; Pandolfi et al. 2003; Costanza et al. 2007). Some of the most fragile may disappear within our lifetimes (Hoegh-Guldberg et al. 2007). As the human population and its influence continue to expand, the demands placed on already over-exploited ecosystems will only become more intense (Dietz et al. 2003; Wackernagel & Rees 2013). Fossil fuels will become scarcer and more expensive, and people will go to ever greater and more destructive lengths to obtain them. As changes in climate cause droughts and famines, nations will go to war to support the needs of their citizens. These issues, and others, are as likely to arise locally and nationally as globally. Prudence in the face of a future likely to be rife with global political insecurity should prompt the global community as a whole to pursue means of sustainably producing and distributing food (Rosegrant & Cline 2003; Devereux & Edwards 2004) and aggressively limiting pollution and destruction of environmental resources (Arora et al. 2011), and the United Kingdom specifically to place a greater emphasis on growing a substantially larger proportion of its own food supplies, timber and biofuel crops (DEFRA 2009). These measures will also conflict with an increasing need for building space and the infrastructure of modern society. Put simply, more must be done with less.

Simultaneously, it is our desire, and our responsibility, to mitigate and correct the damage done to the environment by centuries of industry - not just to secure the health and future of our descendents, but also for the sake of every other species that enables and ennobles our continued existence.

The task of balancing the immediate needs of people against the long-term health of the natural world that we all ultimately depend upon is formidable to say the least, and will surely require all the wisdom humanity can muster to result in a favourable outcome, if indeed such a balancing act is possible at all. This wisdom simply cannot be attained without a solid base of ecological knowledge with which to build upon. It is the aim of this project to contribute towards this base of knowledge by producing tools that will provide us with a greater understanding of the process by which populations of organisms become adapted to the specific, changing environments they find themselves in, known as local adaptation. A better understanding of this process will help us to better understand where our efforts may be most constructively applied.

1.2 THE DAWN OF WHOLE GENOME GENETICS

1.2.1 THE FIRST GENOME SEQUENCES

The revolution in biological information has relied in large part upon the discovery of experimental methods that enable us to directly examine the genetic material of living organisms. Over the last 40 years, biologists have mastered the various techniques collectively known as 'sequencing' that allow us to read off the sequences of molecular building blocks that constitute the DNA of every living thing.

At first, this information arrived in a trickle. Pioneering scientists such as Nobel Prize-winner Frederick Sanger painstakingly reconstructed the diminutive genomes of viruses (Sanger et al. 1977) and bacteria (Blattner 1997) from the results of sequencing reactions applied to short, overlapping fragments of DNA; a jigsaw puzzle strung out in one dimension. For the first time, it became possible to examine the base templates that, through the molecular processes of transcription and translation, direct the formation of proteins from amino acid monomers and therefore, ultimately, the characteristics of every living organism that contribute to their survival or demise (Crick 1958).

This early progress in molecular-based information has already provided enormous benefits for human beings. The knowledge gained from bacterial sequences contributed to programs of genetic engineering in which, by manipulating virus vectors to splice novel genes into bacterial genomes (Cohen & Chang 1973), previously scarce, expensive, or simply unattainable proteins (Goeddel et al. 1979) and drugs (Georgia 2004) could be produced in bulk. This technique was used, for example, to produce human insulin for diabetics in unheard-of quantities. Mass production of proteins and other biomolecules through the culture of genetically transformed microorganisms has since become a substantial industry, and has contributed significantly to human health.

1.2.2 REFINEMENT AND INTEGRATION

As time went on, ever more rapid and accurate sequencing methods based on a diverse range of novel chemistries (Sanger et al. 1977; Ronaghi et al. 1996; Bentley et al. 2008) were developed, and the amount of sequence data on record began to grow exponentially (Stein 2010). To handle this rising flood of data, databases of sequence information were created, storing the genetic sequences of a steadily increasing number of genes from an ever-broader range of organisms. The NCBI (McEntyre et al. 2007) and EMBL (Kanz et al. 2005) databases are the most well known. Accessing the data stored in these databases required the development of software tools (most notably BLAST (Altschul et al. 1990)), which allow scientists to quickly sift and search this large amount of information. Other databases of associated information – for example, protein structure information (Altschul 1997), RNA sequence data (Griffiths-Jones 2003), gene expression data (Edgar 2002), and many more – were also shortly set up as the need arose.

Crucially, many of these databases are now integrated into a single framework of knowledge; a query submitted to one database will also return links to relevant information from associated databases (Sujansky 2001). When data is integrated into a single framework in this way, deeper insights into the underlying processes that shape living things may be attained – as is discussed in later parts of this chapter, and as is demonstrated by the synthesis of genomic and geographic data in the research that constitutes this project. For example, a synthesis of DNA sequence, protein structure, RNA sequence and gene expression data allows us to construct a theoretical model that describes how molecular responses to specific stimuli are effected and regulated, and makes predictions of the results of altering those stimuli on an organism's phenotype which may be tested via further experiments.

However, with the exception of prokaryotic species – whose small genomes could be more readily sequenced in their entirety – the integration of genotypic

data into this framework was restricted to either the reference genomes constructed by sequencing projects, or to experiments targeting specific genes for sequencing in order to search for variation that might explain specific phenotypic differences. While investigation of the genetic basis of specific traits in terms of a small number of variants has been an effective means of gaining understanding since Gregor Mendel carried out his famous experiments (see (Hasan 2004)), the re-integration of statistical modes of thought into genetics demanded new means of cataloguing variation across the whole genomes of many individuals (see Chapter **1.4**). Analyses of this type would remain unfeasible until high-throughput genotyping technology became mature, but the need for them was anticipated (The International HapMap Consortium 2005).

Soon, the base of knowledge, the technology and the infrastructure for sharing data had improved to the point that sequencing projects could attempt to tackle more ambitious targets. While utilising the same fundamental principle of stitching together short fragments of sequence derived from a sequencing reaction as the earliest genomic projects (a process known as assembly (Earl et al. 2011)), newer projects used computers and machines to automate much of the experimental process. Despite advances in speed, reliability and cost over sequencing by hand, a sequencing project still represented a significant investment of time, resources, money and expertise. The *Arabidopsis thaliana* genome project was completed in the year 2000, and marked the first complete genome sequence of any plant (Arabidopsis Genome Initiative 2000).

This initially restricted whole-genome sequencing projects to those eukaryotic species in which the possession of a reference genome would best facilitate the consolidation of knowledge regarding genotypic variation, and the expansion of our understanding of the effects of that variation on characteristics of interest. Sequencing programs were at first therefore directed exclusively – and, to this day, are still directed largely – at 'model organisms': species that have been

chosen for their ease of culture in laboratory conditions and utility in experiments, and for their ability to act as informative models, allowing us to draw conclusions which may be extrapolated with some confidence to similar species (Bancroft 2000). An ideal model species produces a large number of offspring in a short amount of time under safe and easily replicable lab conditions, is commonly found in the wild over a large geographic range, and naturally possesses substantial genotypic variation across traits of scientific and/or commercial interest.

Arabidopsis thaliana fulfils all of these requirements. Its habitable range extends from Scandinavia to North Africa in latitude, and across Eurasia and North America in longitude. It undergoes an annual life cycle, and reaches full maturity in a matter of 6-7 weeks (Boyes & Zayed 2001). It requires relatively little space to grow – a characteristic which lends itself to efficient biological replication of experiments. Upon reaching maturity, a single plant sets hundreds of seeds, which may be easily threshed, stored and sown. Being a member of the Brassicaceae, it is closely related to numerous crop species. Its genome is relatively small, at 137 megabases (Greilhuber et al. 2006).

1.2.3 EXPECTATIONS FOR AND RESULTS FROM THE HUMAN GENOME PROJECT

Around the turn of the millennium, the reconstruction of a species' entire genomic sequence was expected to provide complete answers to long-standing questions. Prior to the Human Genome Project (HGP), the extent of our ignorance of genomics was profound (for review, see (Lander 2011)). The number of protein-coding genes in the human genome could only be guessed at. A few thousand genetic markers were known – enough to create a rough linkage map assigning some heritable disease phenotypes to genes – but the genes underlying the majority of disease phenotypes remained elusive. Analysis of human population history predominantly focused on discerning the histories of single markers such as mitochondrial genes rather than the aggregate whole.

A draft of the human genome was released to the public in 2001 (Olivier et al. 2001). Many – particularly those involved in the non-specialist press – predicted a dramatic, near-instantaneous leap in our understanding of the genetic basis of the phenotype. As the following decade has shown, the data did not bear out such optimistic hopes. However the conclusions drawn from the sequencing projects in general, and from the pioneering HGP in particular, are no less fascinating for it; and the techniques developed and later mastered during the earlier eras of genomic science have often provided invaluable springboards for research on other species and in other fields.

Before the completion of the HGP, the number of protein-coding genes in the human genome was essentially unknown. That number was discovered to be substantially lower than most estimates: 20,000 to 30,000 (Clamp et al. 2007) as opposed to the hundreds of thousands often speculated to exist. But this was just the beginning of the new insights the field of genomics would bring us. It was soon realised that the majority of the genome is constantly being transcribed to RNA (Yelin et al. 2003; Johnson et al. 2005), that this transcriptional turnover is a key component of the regulation of gene expression (Bejerano et al. 2004; Maston et al. 2006), and that a great deal of the evolution of species is more often focused around changes to this regulatory machinery rather than functional mutations in genes themselves (King & Wilson 1975).

Likewise, the technology and methodology of whole-genome sequencing has advanced at a rapid pace in the last two decades, with both the volume of genotype data produced in a given length of time and the number of bases sequenced for a given cost sometimes doubling in less than a year (Stein 2010).

The Sanger sequencing technique was the workhorse of the HGP (and other, earlier genome sequencing projects) but as the HGP was ongoing, much more rapid and cost-effective next-generation sequencing techniques were under development. In order to best ensure the accuracy of the final assembly, the

HGP program planned to obtain at least 10 replicate sequencing reads at any given locus. This depth of coverage was selected to ensure that discrepancies in reported sequences that could be attributed to sequencing errors would be readily apparent, and could be corrected during assembly.

Second-generation techniques, including the Illumina method (Bentley et al. 2008), rely upon a large number of short sequence reads being produced and read in parallel. The individual sequence reads in second-generation sequencing are shorter than those produced via Sanger sequencing (100-250 bases for Illumina vs. routinely up to 1000 bases for Sanger) , but the inherently parallelised process of second-gen sequencing approaches means that a high degree of coverage can be consistently achieved across the genome without undergoing the laborious process of repeated Sanger sequencing (Quail et al. 2012).

While second-generation sequencing methods were first applied to the problem of constructing reference genome sequences, the technology readily scales to obtaining genotypic data from individual organisms, and so became the primary technology employed by modern resequencing projects, including the Arabidopsis 1001 Genomes Project (see Chapter **1.6.1**) (Weigel & Mott 2009).

1.2.4 CATALOGUING VARIATION: THE HUMAN HAPMAP PROJECT

Although the knowledge gained from genome sequencing projects was unquestionably of scientific value, this type of data was unsuited towards actually carrying out genetic experiments. The genome projects produced essentially a catalogue of genes in the human genome. In order to pursue further knowledge – to gain further insight into a species' demographic past, or to investigate the selective pressures still acting upon our population, or to explore the phenotypic consequences of genotypic variation – a catalogue of genetic variation was needed (Manolio et al. 2009). Following on from the completion of the HGP, this type of data has appeared in two phases: firstly in the form of the International HapMap Project (The International HapMap

Consortium 2005a), which recorded allelic variation at the sites of single-nucleotide point mutations (also known as single nucleotide polymorphisms, or SNPs) across 269 human individuals, and later in the form of the ongoing 1000 Genomes Project (The 1000 Genomes Project Consortium 2010), which aims to completely resequence the genomes of at least a thousand individuals and, in doing so, capture all genotypic variation between sampled individuals. Research utilising *A. thaliana* has followed a similar pattern of mass variant identification and resequencing projects, which will be discussed in greater depth in Chapter **1.6.1**.

The main aim of the Human HapMap Project (The International HapMap Consortium 2005) was to investigate the genetic basis of common disease-associated and disease-related phenotypes, such as the progression of particular ailments (for examples, see (Manolio et al. 2008)). Ideally, an investigation of this kind would use whole genome sequence from a large number of individuals (precisely as the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010) is now doing); however at the time, the unfeasible cost of such an undertaking, and the availability of suitable microarray-based assay methods, led the HapMap Consortium to examine SNP markers as representatives of variation within the human genome (The International HapMap Consortium 2005a). The key technology of this project was the 'SNP chip' (Tsuchihashi & Dracopoli 2002): a variant on standard microarray technology (Schena et al. 1995), in which the hybridising probes are formed from genomic sequences corresponding to specific SNP alleles and their flanking sequences.

Ultimately, the Human HapMap Project characterised more than 3 million SNP alleles across 269 individuals strategically selected to represent the breadth of human genetic diversity. A whole-genome map of linkage disequilibrium was then assembled from this data, elucidating the (surprisingly simple) haplotype structure of the human genome (The International HapMap Consortium

2005a), and granting fresh insight into the genetic consequences of events in our species' past, such as a severe population bottleneck tens of thousands of years ago (Gathorne-Hardy & Harcourt-Smith, 2003).

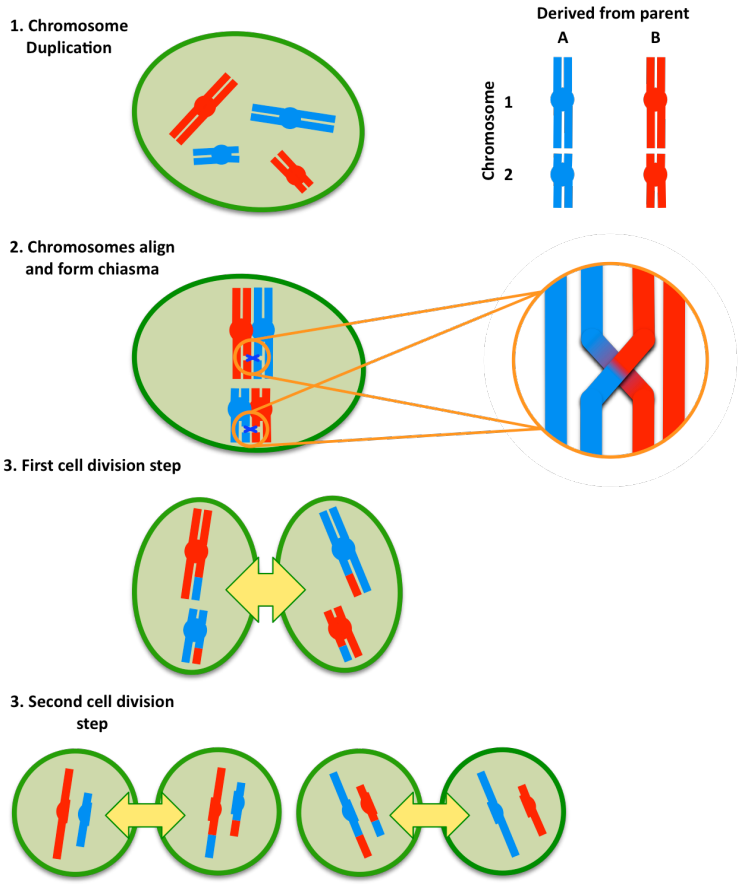


Figure 1 Meiosis and recombination Meiosis is the cell division process by which cells specialised for the purpose of sexual reproduction (gametes) are formed. Four gametes are produced through a replication step and two successive cell division steps, resulting in each gamete possessing the haploid number of chromosomes. To complete the process of sexual reproduction, gametes later fuse, forming a new individual and restoring the diploid number of chromosomes. Diploid parental cells first undergo a replication step (**Step 1**), in which their chromosomes are duplicated. The two replicated chromosomes (chromatids) remain physically attached at the centromeric region, represented here as a circle. In order that each of the two daughter cells resulting from the first cell division step receives the correct complement of chromosomes, the duplicated chromosomes are bound together by protein complexes and by the formation of chiasmate junctions (**Step 2**), and become physically separated prior to the first cell division (**Step 3**). When this separation occurs, the chiasmate junctions between homologous chromosomes remain intact, effecting an exchange of genetic material between the two chromosomes; note the distribution of genetic material in chromatids following the first cell division. Replicated chromosomes separate before the second cell division (**Step 4**). An important consequence of this recombinatory process is that over generations, sets of alleles become mixed together. If chromosome B1 is viewed as a new haplotype, for instance, that haplotype must be considered to be shorter in the gametes possessing chromosomes which have undergone recombination, since some of that haplotype's alleles have been supplanted by a different set of alleles originating from the other, homologous chromosome. See Chapter **4.1.2** for discussion of the consequences of this routine exchange of genotypes between chromosomes on a population in terms of haplotypes.

1.2.5 KEY GENETIC CONCEPTS AND METHODS

In order to fully explain the findings of the HapMap project and the further possibilities it opened up, it is necessary at this point to explain some basic concepts that underpin the science of genetics.

Meiotic recombination is the key phenomenon at the heart of almost all others of relevance to this project, from the origination of an individual genotype, through linkage disequilibrium, to the persistence of haplotypes in a population and the pattern of genotypes dispersed across an entire continent. (See Figure **1** for a visual explanation of meiotic recombination). Meiosis is the cell division process by which gametic cells, possessing a haploid genotype, are produced from parental cells possessing a diploid genotype. In its earlier stages, pairs of parental homologous chromosomes are aligned on a plane in the centre of the cell, in order that the diploid number of chromosomes may be neatly and consistently reduced to the haploid number. This is affected by a large and complex set of molecular machinery that draws the chromosomes of each homologous pair towards opposite ends of the cell prior to the actual division of the cytosol. In order that the paired homologous chromosomes do not separate prematurely (a situation that would most likely result in the two daughter cells receiving an incorrect number of chromosomes – a condition known as aneuploidy), their DNA strands are broken and rejoined to each other at matching points along the length of each chromosome (Morgan 1921). When the two chromosomes are separated towards opposite ends of the cell, the DNA strands remain in their rejoined state, with the consequence that an exchange of homologous genetic material between the pair of chromosomes has taken place. This process is known as recombination; the points at which DNA is broken and rejoined are known as chiasma or crossovers (for a full overview of meiosis and the history of its discovery, see (Schwartz 2009)).

Linkage is the tendency for alleles of different genes that reside on the same chromosome to be inherited together, due to meiotic recombination being less likely to separate any given pair of alleles than to keep them together in their current configuration (Bateson et al. 1909). The probability of recombination breaking the linkage between any given pair of alleles – i.e., of a chiasma forming between the two loci – is proportional to the genetic distance between the two loci. It is therefore possible, by measuring the rates at which recombinant phenotypes emerge in the F_1 and F_2 generational offspring of parents of a known genotype, to measure the relative distances between genes possessing alleles causing particular, observable differences in phenotype. This is the basis of linkage mapping (Griffiths et al. 2000), and constituted some of the earliest genetic experiments (Morgan 1911).

Table 1 Assumptions of Hardy-Weinberg equilibrium

- Diploidy
- Only sexual reproduction
- Non-overlapping generations
- Random mating
- Large/infinite population size
- Allele frequencies are not sex-linked
- No selection
- No migration
- No mutation

Obtaining accurate results from this technique relies upon recording the phenotypes of a large number of offspring – typically, hundreds or thousands (Liu 1997)**.** Undertaking such an experiment via controlled matings with human beings would of course be impractical and unethical , so scientists constructing linkage maps of the human genome have instead relied upon medical and pedigree data to produce linkage maps (Lander & Green 1987). *A. thaliana*, however, can be readily grown in large numbers, and since individuals reach maturity only months after germination, experiments spanning multiple generations are feasible.

Hardy-Weinberg equilibrium describes a simple mathematical model of an idealised population essentially free from any stochastic or external influences (Hardy 2003; Weinberg 1908). The model assumes a number of characteristics of a population, listed in Table **1**. Under the model, neither the frequency of an allele within the population nor its level of heterozygosity change from one generation to the next, provided that all of the assumptions of the model continue to hold true. This means that the power of Hardy-Weinberg equilibrium as a tool of inference is as a null hypothesis applicable to a wide variety of questions; when an extant population is observed to differ from a state of Hardy-Weinberg equilibrium, it may be concluded that at least one of the assumptions of the model does not hold true for that population. Additional knowledge of the population often allows the violation of the model to be deduced; when a significant deviation from Hardy-Weinberg equilibrium is found in autosomic alleles within a large but otherwise isolated population of annual, outcrossing plants, for example, the most likely remaining explanations for the observed deviation are either non-random mating, or that some alleles are favoured over others by selection.

Linkage disequilibrium (LD) is effectively an extension of this concept of linkage (Reich et al. 2001). Under conditions of Hardy-Weinberg equilibrium, linkage may be expected to follow a predictable pattern (Wigginton et al. 2005). Loci at a given genetic distance from each other should, all other factors being equal, show a consistent and predictable ratio of parental and recombinant genotypes in their offspring, controlled by the frequency at which recombination happens within the region of the genome separating them. Should a population show a higher (or lower) frequency of recombinant genotypes between two loci than expected, however, the two loci are in linkage disequilibrium. Since such a finding represents a departure from Hardy-Weinberg equilibrium, it follows that the potential causes of linkage disequilibrium are the same as those that might cause departures from Hardy-Weinberg equilibrium (Deng et al. 2001).

Haplotypes are genetic elements comprising groups of alleles along the length of a portion of a chromosome that are inherited together (The International HapMap Consortium 2005). They are a phenomenon of linkage, and their persistence over time within a population is controlled by the degree of linkage disequilibrium between the loci in question (Wall & Pritchard 2003). Crossing over during meiosis has the effect of breaking a haplotype up, since a different set of alleles are spliced into some of the loci along the haplotype's length (see Figure **1**). When utilising polymorphism data such as SNP data, it is usually taken as a safe assumption that the SNP alleles of a haplotype reliably predict the variation at adjacent loci (where the genotype may not be directly known) (Karp 2003), since alleles at these loci are tightly linked with both the SNP alleles and each other; the probability of two crossover points forming such that two closely-linked SNP alleles remain in their original configuration while any other alleles between them are altered is exceedingly low. This was the justification given for using SNP data in the HapMap Projects (The International HapMap Consortium 2003; Kim et al. 2007).

While linkage-based mapping methods remain a key tool in the geneticist's toolkit – indeed, reliable assembly of genome sequences without the guidance of prior linkage mapping remains a significant challenge (Li et al. 2010) – they are not without limitations. Geneticists have, in recent years, come to recognise that phenotypic traits are often controlled not by alleles at a single locus, but by the summed effects of alleles spread across many genes (and, conversely, that variation in one gene may contribute to variation across multiple phenotypes – a condition known as pleiotropy (Sivakumaran et al. 2011)). (See Chapter **1.4** for further discussion of the background of quantitative genetics). Human studies of common disease phenotypes have shown that this is likely to be the norm (Plomin et al. 2009), and that the single allele-linked phenotypes upon which so much work in the field of genetics was based are, in reality, more often the exception than the rule. Despite advances such as intercross mapping, linkage mapping techniques still struggle in cases where a large number of

genes each exhibit variation which makes only a minor contribution to a phenotype. A genome-wide association study (GWA or GWAS) is better suited to this type of situation (Plomin et al. 2009; Korte & Farlow 2013).

A GWAS typically involves a comparison between two groups of individuals: one 'case' group presenting the phenotype under investigation, and one control group, with a phenotype different from that under investigation (Kruglyak 2008). In the context of a genetic disease susceptibility investigation – a common theme in human-based GWAS – this generally corresponds to a case group of individuals displaying disease or disease-susceptible phenotypes, and a control group of healthy or otherwise disease-resistant individuals (for example, (The Wellcome Trust Case Control Consortium 2007)). The genotype of each participating individual is recorded, usually in the form of SNP allele variation measured using SNP genotyping arrays similar in principle to those used by the HapMap projects. The analysis then comprises a search for alleles found more commonly than expected by chance in the variant group. For each SNP allele, the 'odds ratio' – the ratio of the allele's frequency in the case group to its frequency in the control group – is taken; a chi-square test may then be applied to determine whether the allele's odds ratio is significantly different from all other alleles' odds ratios (most of which will have no effect on the phenotype in question and will therefore provide a good approximation of the range of frequencies within each group which may be expected from pure chance) (Pearson & Manolio 2008). Any allele showing a significant deviation from chance expectations is then said to be 'associated' with the case phenotype. A more advanced version of this analysis might expand the case/control model to find genotypes associated with a range of variation of a quantitative phenotype (for example, (The Wellcome Trust Case Control Consortium 2007)).

When working with phenotypic traits that fall along a continuous scale rather than into discrete categories, the association mapping approach enables

geneticists to quantify the degree to which variation within one gene affects the phenotype as a whole. This view of genetics does not contradict the Mendelian model; individual alleles are still subject to Mendel's laws, though quantitative genetics provides a framework that allows us to understand how a complex, continuous scale of phenotypes may emerge from the interaction of multiple discrete elements (Plomin et al. 2009; Mackay et al. 2009).

In order to reliably estimate the frequency ranges within a group expected purely by chance, however, knowledge of the genome's linkage disequilibrium and the underlying structure of a population must be integrated into the analysis. A failure to account for these factors is likely to lead to false positives; should an allele affecting the trait under investigation happen to be linked to another locus which in reality has no effect on that phenotype, both loci may be incorrectly recognised as associated with the trait (Pearson & Manolio 2008; Astle & Balding 2009). This may be taken into account by applying statistical corrections for known linkage disequilibrium, and by carefully selecting participants in both groups in order that population structure is negated as a confounding factor(Zhao et al. 2007).

Under conditions of Hardy-Weinberg equilibrium, any individual has an equal probability of producing offspring with any other. When a population deviates from this assumption – a species may, for example, inhabit two isolated areas that offer relatively few opportunities for an individual to move from one area to the other – the population is said to be structured (Wright 1950). Since this phenomenon will be discussed at length in Chapter **3** it is enough, for the moment, to say that without carefully accounting for the demographics of a population, it may be difficult to distinguish linkage disequilibrium attributable to natural selection from linkage disequilibrium caused by population structure.

1.2.6 GENOMIC DATA IN THE PRESENT AND THE FUTURE

With these techniques and data providing a foundation of genetic knowledge, it becomes possible to investigate more far-reaching questions regarding the ecology and history of a species. Where did the genotypes we see in a given area today arise from, and why are they found in the places where they are now found? Which environmental conditions or other cohabiting species drive natural selection on a population in a given area, and how does the species respond to those selective pressures? Just as the study of the inheritance of easily observable traits in model species led to the formulation of a theoretical framework of genetics which applies to any sexually reproducing species, an attempt to extend that framework to ecological questions of this kind may advance our understanding to the point that reliable predictions may be made for any comparable ecological circumstance.

While fascinating in their own right, these lines of inquiry also provide us with practical benefits: an intellectual framework within which we may understand the interactions between crop species and pathogens, thus helping agriculture to progress more efficiently and more sustainably; knowledge of the manner in which wild species are likely to respond to a changing climate, and therefore how we might go about protecting them; and advances in methods and understanding which feed almost directly back into research on human genetics and health.

The requirement for and benefits of incorporating genetic knowledge into a wider ecological framework in the specific context of this project are further discussed in Chapter **1.8**.

1.3 UNVEILING POPULATION HISTORIES FROM GENOMIC DATA

1.3.1 DEMOGRAPHY AND POPULATION STRUCTURE: METHODS AND CONCEPTS

A complete understanding of both ecology and evolution relies upon a comprehension of the fundamental properties of a population: the mathematical and statistical characteristics of a group of organisms as they persist through time. The field of population genetics works with this branch of mathematics and biology. Two key phenomena underlie almost all research in population genetics: population structure, and genetic drift.

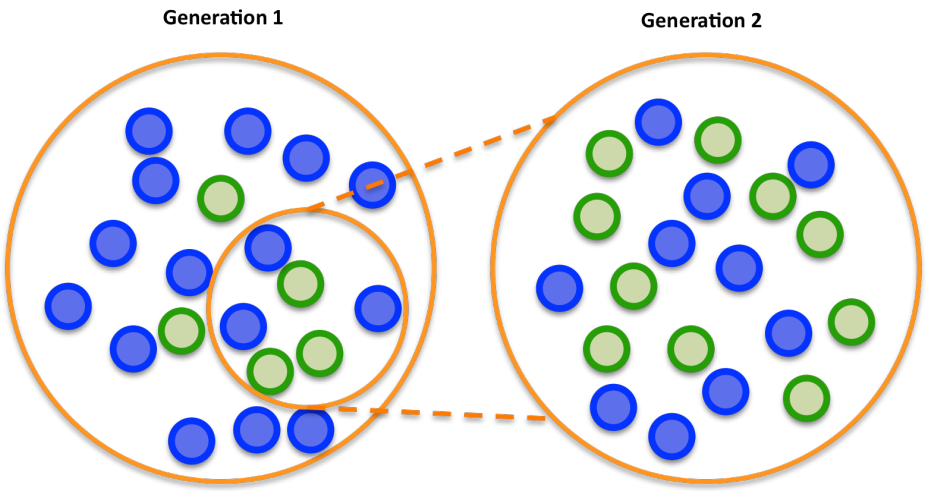


Figure 2 Genetic drift As the number of individuals in a population decreases, sampling error is increasingly likely to become a significant factor determining the frequency of genotypes passing from one generation to the next. This diagram shows how sampling error may cause the frequency of genotypes to substantially change (from a ratio of 3:1 to 1:1) in a single generation. The term 'genetic drift' refers to such changes in allele frequency caused by sampling error. See Chapter **1.3.1** for further discussion of genetic drift.

Genetic drift is a change in the frequency of an allele from one generation to the next caused simply by random sampling (Masel 2011) (see Figure **2**). A common analogy is repeated rolls of a die: over a large to infinite set of repeat rolls, the six possible outcomes will tend toward an equal 1/6 frequency; but over smaller numbers of repeat rolls the difference in observed frequency is likely to differ from the expected result, with smaller numbers of repeats

tending to show a more pronounced departure from expectations. In a biological context, this means that (all else being equal) in large populations an allele's frequency is unlikely to change significantly from one generation to the next, but that in very small populations an allele might rise to fixation or decline to extinction over a small number of generations, due to nothing more than stochastic effects. It is for this reason that Hardy-Weinberg equilibrium (Hardy 2003; Weinberg 1908) assumes a large to infinite population size.

Conversely, the extent of genetic drift may be used to estimate the number of individuals in a population (Wright 1938). If the simplifying assumption that a population follows an entirely random mating model (i.e., with no population structure) is applied, the number of individuals in the population is inversely proportional to the degree of drift between generations.

Stated simply, population structure is a result of non-random mating between individuals within a population – a bias towards a given individual selectively mating with one individual or group over another (Weir & Cockerham 1984; Pritchard et al. 2000). This deceptively simple basis gives rise to far-reaching consequences of huge scientific interest: non-random mating may manifest as sexual selection, for example; or, should a population happen to be divided into several sub-groups with limited interchange between them, there will be a bias towards mating within the sub-group, rather than between sub-groups.

Population structure thus causes deviation from Hardy-Weinberg equilibrium in two distinct ways: each sub-population is subject both to increased drift due to the smaller number of individuals in each group, and migration between the groups also pushes the allele frequencies in each group away from equilibrium. This metapopulation model has formed much of the foundation of the field of population genetics (Hanski 1998), and both provides a means of understanding the theoretical genetic and demographic aspects of some of the key proposed mechanisms of speciation (Levin 1995; Gavrilets et al. 2000; Ramos-Onsins 2004), and generates particularly useful predictions when

applied to the practical goal of conserving fragile and fragmented populations of endangered species (Fahrig & Merriam 1994; Marsh & Trenham 2001).

Drift may have a profound influence on the allelic composition of a population if the number of individuals in that population shows a large degree of variation over time. A population may, in its past, have shrunk to a small number of individuals, and later recovered. When such a 'bottleneck' occurs, the relative proportion of alleles in the gene pool will initially change rapidly between generations, due to the increased effect of random sampling; but when the population recovers, the changes in relative proportion are maintained in the new, larger population. Analysis of genotypic evidence can reveal historical instances of events of this kind (Nordborg et al. 2005). The same principle may apply should a small group of individuals become reproductively isolated from the wider population and go on to establish a second population. In this case, the bottleneck effect is known as the 'founder effect', and is also readily identifiable from analysis of allele frequencies (Nordborg et al. 2002).

1.3.2 HISTORICAL POPULATION GENETICS FROM HUMAN DATA

Genomic data has, in recent years, granted new insight into the history of *Homo sapiens*. Analyses of the present distributions of human genotypes have revealed details of the migratory history of our ancestors and their evolutionary cousins as they arose in Africa and spread, in multiple waves, across every other habitable continent (Gunz et al. 2009). Other analyses have shown evidence of a time in the evolutionarily recent past when *H. sapiens* was pressed almost to extinction, passing through a narrow population bottleneck (Gathorne-Hardy & Harcourt-Smith 2003). Recent analyses comparing the human and chimpanzee genomes have also been used to investigate the possibility of inter-crosses between hominid species in the more distant past (Patterson et al. 2006). Further development of knowledge in the field of human population genetics was one of the major stated goals of the Human HapMap Project.

Knowledge of population genetics has also proven key to understanding selection acting upon human populations, such as selection favouring alleles conferring resistance to highly prevalent and deadly diseases (Sabeti et al. 2002).

1.3.3 POPULATION GENETICS FROM ARABIDOPSIS THALIANA DATA

Due to the manner in which *Arabidopsis thaliana* is often encountered in the wild – often in sites extensively disturbed by humans, including gardens, roadsides, agricultural wall sites and railways (Mitchell-Olds 2001) – along with its small, light seeds that potentially lend themselves to inadvertent dispersal by human movement, some had supposed admixture across the species' native range to be extensive enough that population structure to be practically non-existent, on account of rapid and recent population expansion(Innan et al. 1997; Innan & Stephan 2000). Another, competing hypothesis proposed that *A. thaliana* did indeed exhibit population structure, reflecting a post-Pleistocene expansion of the population across Eurasia (Sharbel et al. 2000). Under this hypothesis, it may be expected that the wild population possessed a structure reminiscent of a distinct branching pattern, resulting from the expansion of the population from the site at which the species first arose into new, previously uninhabited regions. This expectation was not borne out by evidence (Nordborg et al. 2005), indicating a high degree of admixture; however, numerous investigations have reported distinct population structure.

Sharbel *et al.* (Sharbel et al. 2000) reported results from AFLP analysis that appeared to indicate that the wild *A. thaliana* population follows the 'isolation by distance' (IBD) model. Under this model, genetic similarity of individuals is inversely correlated with the distance separating their home ranges or growth sites, and populations are expected to form a 'stepping stone' arrangement, in which populations become established in small habitable sites separated by expanses of uninhabitable (or transiently inhabitable) space, occasionally dispersing seeds from one to the next (Kimura & Weiss 1964; Hardy &

Vekemans 1999) . Under conditions of low outcrossing and geographic range much larger than typical dispersal distance, the eventual emergence of isolation by distance was proposed to be essentially inevitable, provided that the habitable range of the population is not simply restricted to several highly isolated sites.

Haplotype analysis of 10 loci assayed from samples gathered across Eurasia by Beck *et al.* (Beck et al. 2008) again suggested the existence of an isolation by distance population structure, and also showed an east-west split demographic split which Beck *et al.* (2008) attributed to population dynamics consistent with those from other species inhabiting ranges affected by Pleistocene ice sheet movements (see Chapter **3.1.3** for a more in-depth examination of this population structure).

Platt *et al.* (Platt et al. 2010) later applied a QT-clustering algorithm to a large-scale set of genomic data (a precursor to that released by the Arabidopsis HapMap project) containing the results of an assay of 149 SNPs across ~6000 wild *A. thaliana* accessions, including many from previously unsampled regions. This analysis aimed to investigate the ancestry of accessions in this dataset relative to different geographic scales (local, regional and continental), and to estimate the rate of outcrossing that occurs in nature. From this, they concluded that in the wild, *A. thaliana* outcrosses only rarely (97% selfing rate, on average), and that the population structure follows the isolation by distance model. This paper raised the possibility of using genome-wide polymorphism data to measure the effective population size (the number of organisms in an ideal population – that is, under an ideal model of perfectly random mating – that would cause the observed degree of genetic drift) of *A. thaliana*, but ultimately declined to make an attempt at doing so.

Additionally, it was proposed by Platt *et al.* (2010) that isolation by distance emerged in a newly established population only over a period of time and after several waves of migrants arrive, since a population derived from a small

number of founders offers few opportunities for new haplotypes to arise (and thus for individuals in the population to become differentiated) through the action of meiotic recombination upon diverse genotypes. Platt's observations from recently established populations (i.e., those in the USA) showed a highly variable, inconsistent distribution of genetic similarity across distance classes, which developed to follow the trend of isolation by distance much more closely and consistently in more well-established populations.

While Platt *et al.* (2010) and Beck *et al.* (2008) had examined population structure at a continental scale, Bomblies *et al.* (Bomblies et al. 2010) examined the structure of *A. thaliana* populations on smaller scales: that of local stands and individual plants. As with previous findings, Bomblies *et al.* (2010) reported significant differentiation in genetic variation between sites, and a general pattern of isolation by distance.

These claims deserve another, independent examination using different methods and more dense genotype variation data. Should this model of population structure indeed prove to hold true in the denser polymorphism dataset used in this project, the model will provide a sound means of generating null hypotheses in the investigation of evolutionary and more advanced demographic phenomena. Chapter **1.6** discusses the advantages of proposing and testing hypotheses of this kind.

It is also important to remember that the isolation by distance model predicts a trend, rather than an absolute rule. While this trend may act as an excellent general model for the structure of the *A. thaliana* population, it resolves little detail regarding specific instances of gene flow in the history of the species. More complex demographic and population genetic analyses are required to detect specific migratory and selective events in the recent history of the species. Occasionally – particularly as human influence over the ecology of Europe has so drastically increased in recent centuries – gene flow deviating from this model may occur. Long-distance dispersal of small numbers of

individuals across large distances may not exert any significant effect on the observed population structure, since the migrant haplotype is likely to be rapidly broken apart by meiotic recombination as it disperses through the immediate population. Larger invasive or migratory events, though, may be expected to produce clear deviations from the predicted structure, amounting to unusually extensive genotypic similarity between samples collected at distant sites. Chapter **3.3.1** of this project is dedicated to the examination of potential examples of larger-scale migration events.

1.4 QUANTITATIVE GENETICS: REVEALING THE GENES BEHIND THE TRAITS

1.4.1 TOWARDS A BETTER UNDERSTANDING OF THE CAUSES OF COMPLEX PHENOTYPES

In order to gain anything approaching a complete understanding of the relationship between the dynamics of alleles within a population and the physical traits upon which selection acts, it is necessary to understand how an individual organism's phenotype is created and controlled by the genotype behind it.

In the last decade or so, our understanding of quantitative genetics has increased markedly. Work in the field of human genetics – particularly the genetics of common heritable diseases – has frequently driven advances in understanding and improvements in methodology in this area (Plomin et al. 2009; Mackay et al. 2009). For much of its history, the field of genetics was divided into two camps, with relatively little interaction between them: "Mendelians", who sought to explain the descent of observed traits over generations in terms of molecular inheritance; and "biometricians", who despaired at finding simple Mendelian explanations for complex traits, and instead sought quantitative explanations for the phenomena they observed. As Plomin (Plomin et al. 2009) points out when discussing how this divide was resolved into the modern framework of quantitative genetics, both camps were partially correct and partially wrong.

Although a theoretical reconciliation between these two branches of genetics was proposed by R.A. Fisher as early as 1919, in which the overall phenotype is constructed from the summed phenotypic effects of alleles at multiple loci which each independently obey Mendel's laws (Fisher 2012), biometricians and Mendelians generally remained separated by their very different goals for some considerable time. Mendelians continued to seek out genes – and alleles – explaining variation in particular traits, while biometricians instead measured

the extent to which genetic factors (as opposed to environmental or developmental factors) contributed to variation in traits. A practical synthesis of the two disciplines only became truly feasible with the advent of high-density polymorphism data (of precisely the type utilised in the HapMap Projects and in this project) and the GWAS.

As previously mentioned in Chapter **1.2.5**, genome-wide association studies have proven to be powerful tools in studying the relationship between genotype and phenotype, provided that care is taken to eliminate any associations between alleles and sample groups arising from population structure or sampling bias. One of the major conclusions arising from these studies is that, as the Mendelians predicted, complex traits are typically controlled by a number of genes, with variation in each generally conferring a small change in the overall phenotype (and conversely – and more surprisingly – a substantial degree of pleiotropy, in which variation in one gene affects multiple phenotypes (Sivakumaran et al. 2011). Another major conclusion is that, much as the biometricians proposed, complex traits and the genotypes responsible for them may be effectively discussed in quantitative terms – both in terms of the degree to which any given allele affects the genotype, and in terms of the measurement of the trait itself. This is a crucial point in the study of genetic susceptibility to, and resistance against, disease: it is often tempting to think in simple terms of outright vulnerability or resistance, but resistance/susceptibility traits observed in real organisms more often fall along a somewhat less intuitive quantitative scale. See Chapter **4.3.1** and Figure **22** for an example of a complex, pseudo-quantitative resistance response in *A. thaliana* to attack by an oomycete pathogen. Other examples are also present in the literature (Kover & Schaal 2002)

Informative as it is, the GWAS approach can be constrained by confounding factors (including population structure) requiring additional statistical correction and care in selecting unbiased sample groups. In order to truly

confirm a GWAS finding beyond reasonable doubt, it is necessary to relate the results back to genetic experiments carried out with real organisms – to demonstrate that variation at the loci identified by the GWAS really does cause variation in the trait in question. Fortunately, modern knowledge of quantitative genetics has enabled a powerful and effective means of doing exactly that, which combines some of the best features of linkage and association mapping.

The technique of mapping quantitative trait loci (QTLs) using Recombinant Inbred Lines (RILs) has granted researchers another way to identify parts of the genome associated with phenotypes. This approach has been applied to a considerable number of traits in other model organisms, though of course it cannot be applied in human-based studies (for the same reasons as those discussed in Chapter **1.2.5** for traditional linkage mapping) unless an already extant lineage with a known pedigree is sampled and appropriate statistical corrections applied.

In a typical non-human-based QTL mapping study, a small number of individuals - at least two are required, but more may be used – are intercrossed for at least two generations. Following this, the offspring may then isolated and self-crossed for several more generations in order to create a large number of RILs, or may be phenotyped directly (Doerge 2002). Due to the meiotic recombination (Figure **1**) occurring at each generation of the repeated intercrosses between lines, each line exhibits a patchwork of the genotypes from the parental generation. When the lines are then isolated and self-crossed, the genotypes of each line remain stable over multiple generations, since the individuals within each line are essentially genetically identical. Each line is genotyped using a SNP microarray or simple PCR-based assay, in order that any given locus in each line's patchwork of genotypes can be attributed to having arisen from one of the parent individuals.

The phenotype of each RIL is then observed and recorded, and cross-referenced against the record of the genetic patchwork uncovered by the SNP assays. At most loci, the pattern of SNP markers will not align well against the pattern of phenotypes; however, at loci possessing genetic variation responsible for changes in the observed trait, the pattern of SNP markers will match the pattern of phenotypes to a statistically significant degree. This is represented graphically in Figure **3**.

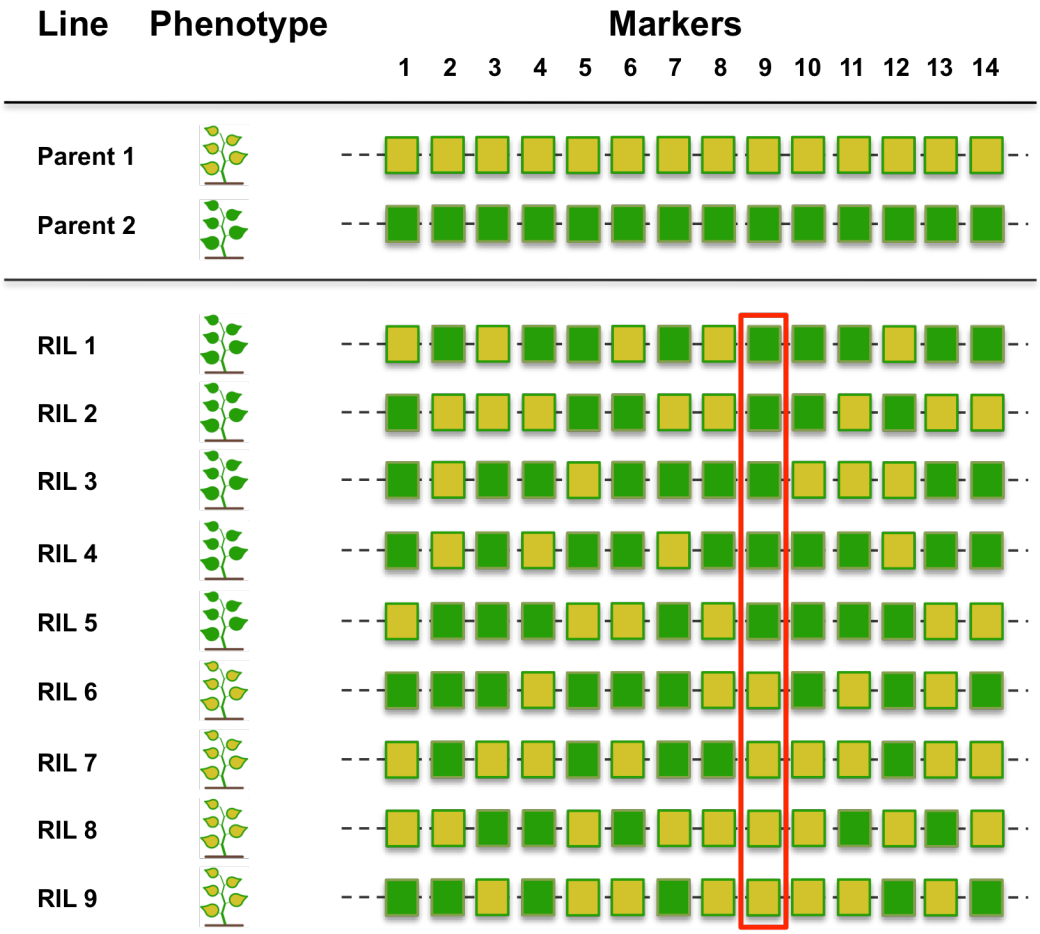


Figure 3 QTL Mapping Process In a typical quantitative trait locus mapping experiment, two parent plants possessing different phenotypes (represented here by leaf colour) are caused to mate. The F1 offspring are intercrossed with each other. The F2 offspring are either self-fertilised or crossed with siblings for several generations to produce recombinant inbred lines (RILs). Genetic markers (typically SNPs) are then assayed in each RIL. Markers displaying a statistically significant co-segregation with the phenotype under investigation (indicated by the RED box) correspond approximately to loci at which genotypic variation induces the observed changes in phenotype.

In practical terms, this technique gives plant biologists a relatively quick and inexpensive means of examining an organism's entire genome for genes possessing alleles that control aspects of the organism's phenotype – and, unlike traditional linkage mapping, may be used in cases where more than one gene contributes to a phenotype. Since this technique also establishes a population along a known pedigree this method is, like linkage mapping, also not susceptible to bias arising from population structure, though it is susceptible to biases arising from unequal distribution of crossovers during meiosis, or non-random segregation of chromosomes (Doerge 2002).

A similar technique also allows extremely fine mapping once the rough genomic locations of variation quantitatively linked with a trait are known: near-isogenic lines (NILs), possessing genetic variation at only a very few loci within the candidate region, may be created (Keurentjes et al. 2007). NILs are subjected to a similar analysis as RILs, though the more scarce nature of the genotypic variation allows the loci possessing variation affecting the trait to be identified more precisely.

While the technology and methods for identifying the causative genes behind any given phenotype have considerably improved in recent years, our actual knowledge of how phenotypes arise from genotypes remains an area of ongoing research, with many questions still unanswered and details still awaiting investigation. Any eukaryotic organism possesses many thousands of genes, and even in a species as comprehensively investigated as *A. thaliana*, the functions of a substantial fraction of those genes remain unknown, or have been only putatively identified through evidence of up- or down-regulation in response to certain stimuli in microarray experiments (for example, (Chen et al. 2002; Goda et al. 2002; Popescu et al. 2009)).

One of the aims of this project is to contribute to the continuing accumulation of knowledge in the areas discussed in this section.

1.5 UNDERSTANDING ADAPTATION FROM GENOME-WIDE DATA

1.5.1 OPEN QUESTIONS IN ECOLOGICAL GENETICS

Accurate, quantitative means of mapping the genotypes responsible for phenotypes has had – as the previous chapter discussed – a revolutionary impact on our understanding of genetics, and the benefits of this new level of understanding are currently also filtering down through more applied fields as diverse as proteomics, food security, and medicine. However, the quantitative model of genetics, together with the modern understanding of population genetics, also grants the possibility of considerable new insight into the realm of ecology: the study of the relationship between organisms and their environment.

Genomic data in combination with geographic data allows us to study a breadth of subjects: the demographic histories of present-day populations, tracing their spread and migration over thousands of years (François et al. 2008); the impacts of human action upon wild species (Platt et al. 2010; Beck et al. 2008) and the genotypes controlling traits of ecological importance – traits responsible for a species' ability to fill an ecological niche – and how selection has acted upon those genotypes to produce their current state (Bergelson & Roux 2010). This latter subject is the major focus of the field of ecological genetics. Unsurprisingly, given the recent advent of large-scale genomic analysis, this is a field where the power of genomic data is only recently beginning to be brought to bear.

Previous research in the field of ecological genetics has built up frameworks of knowledge contributing much to our current understanding of population genetics and ecology, but has also opened up many questions still in need of answers. Some of the open questions in ecological genetics of greatest relevance to this project are discussed below.

- Which genes confer ecologically important variation in phenotype, and why does variation exist within these genes?

Feder and Mitchell-Olds (Mitchell-Olds et al. 2008) state that it has "long been one of the main goals of the fields of ecological and evolutionary genetics to understand genetic basis of traits of ecological significance – those which affect an organism's fitness in the wild". That remains as true today as ever. Although research on many facets of this broad question has been underway for decades, most of the details regarding the genes influencing variation responsible for changes in fitness in specific habitats remain to be filled in.

- Does disease resistance evolve primarily through series of selective sweeps, or through balancing selection?

A long-standing question centres on the means by which disease resistance (discussed in the next chapter as a highly informative set of case studies for ecological genetics and evolutionary biology) evolves in response to the constantly shifting challenges of parasite species that are themselves rapidly evolving. Selective sweeps occur when novel, advantageous alleles arise through mutation and are rapidly driven to fixation, 'sweeping' other, linked alleles to fixation through the hitchhiking effect (Smith & Haigh 1974). Balancing selection, conversely, results instead in an equilibrium of genotypes; a commonly expressed example is that of frequency-dependent selection, where an allele is favoured by selection simply because it is rare. Unlike the newly arisen alleles favoured in a selective sweeps, alleles may, in theory, be maintained in the equilibrium of balancing selection for a very substantial length of time. Examples of both selective sweeps and balancing selection have been found in the context of disease resistance, but it remains unknown which of the two models is the more common.

- How do developmental responses to environmental conditions affect fitness?

As discussed in the context of flowering time, evidence seems to indicate that developmental characteristics play an important part in determining the suitability of an organism to its environment. The exact details of the relationships between developmental features and fitness to given environments are, as with the majority of other phenotypes, only beginning to seriously be explored.

Although this project was focused primarily on the interactions between organisms and their environment in the context of host/pathogen conflicts, it also presented an opportunity to examine evolutionary responses, including those in development-associated genes, to abiotic conditions across several habitat types.

- How does evolutionary change differ in response to varying speeds and intensities of changes in selective pressures?

Lab experiments with populations of single-celled organisms have shown that evolution proceeds in a distinctly different manner in cases where selection pressures alter gradually over time to cases in which selection pressures alter more rapidly (Collins & Meaux 2009). Specifically, rapid changes in selection pressures initially favour mutations of large effect on the phenotype in question and later favour mutations of small effect as the phenotype moves toward a more optimal point in its parameter space, whereas more gradual changes in selection pressures consistently favour mutations of small effect throughout the course of the alteration in pressure. Notably, Collins and de Meaux found that the latter situation may induce overall greater fitness to new environmental conditions than the former.

Given that mutations of small effect may be more likely to arise in a phenotype controlled by many genes (since even a mutation of large effect in one gene is then likely to exhibit only a small alteration of the overall phenotype), it is reasonable to suppose that more gradual changes in ecological conditions will

drive selection to act more commonly on traits governed by quantitative genetics, and vice versa; and, consequently, that the number of genes associated with a trait exhibiting signatures of selection may likewise reveal something of the nature of the pressure driving their selection. This has been little explored in wild populations, however – a scarcity of knowledge that this project partially aims to address.

These questions highlight the close relationship between ecology and evolution; it is practically impossible to attain any meaningful understanding of one without taking the other into account. It is therefore necessary, at this point, to review key concepts of evolutionary and ecological theory.

1.5.2 EVOLUTIONARY CONCEPTS

At the very heart of evolutionary theory is the concept of fitness: the idea that some individuals within a population possess genotypes that ultimately make them better able to survive, out-compete rivals and reproduce in a given set of environmental conditions than other individuals with different genotypes. When this disparity between genotypes manifests in a population, it is likely that the following generation will possess genotypes at a different frequency to that of their parents; some alleles – those better suited to the current environment – will tend to be more common (these alleles are said to be favoured by selection), and vice versa. Such an occurrence is, of course, recognisable as a departure from Hardy-Weinberg equilibrium, and may be quantified by the degree of that departure from the null hypothesis of neutral equilibrium. The implication of this is that, over time, the gene pool of the group of organisms occupying a given habitat will tend to become filled with alleles that produce phenotypes well adapted to that habitat (assuming no other departures from Hardy-Weinberg equilibrium are in effect), and conversely that alleles producing phenotypes less well adapted to that habitat will become less common and eventually disappear altogether. In the context of most ecological literature, a population thus adapted to a habitat is referred to as an

'ecotype', though in *A. thaliana* literature this term is used interchangeably with the term 'accession' (Mitchell-Olds 2001). Characteristics of the environment that cause one allele to be favoured over another – i.e., causes that reduce the fitness of genotypes – are referred to as 'selection pressures'.

This seems simple enough at first glance; yet while some relatively obvious examples of differential fitness attributable to specific genotypic variation have been discovered in the 150 years since Darwin first put forth his theory of evolution by natural selection, in most cases definitively linking genetic variation to fitness has proved laborious at best and frustratingly difficult at worst, even with the increasing provision of large-scale analyses and datasets. Conclusively linking the allele causing sickle cell anaemia in humans to resistance against infection by *Plasmodium* species, for example, required decades of research (Robert et al. 1996), and investigation continues to this day on the exact causes of this interaction of phenotypes.

A more profound understanding of selection, built up over the subsequent decades, has since come to allow us to understand these more difficult cases; though as the open questions described above show, this is still an ongoing area of much active research. Just as modern quantitative genetics makes it possible to identify and understand the contributions a set of genes makes to a physical trait, modern practices of experimental design in ecological experiments are making it possible to detect and examine the process of selection in a much broader range of contexts. Methods of doing this are discussed in greater detail in Chapter **4.1.3**.

1.5.3 FITNESS

In order to continue this line of research, it is also necessary to understand several phenomena relating to evolutionary theory. The existence of these phenomena has been known for decades, though research in ecological genetics has only recently begun to probe them in a comprehensive manner.

When two species are in competition with each other for the same finite resource, there is a selective pressure for each to collect or control as much of the resource as possible, which has the consequence that the resource is denied to the other. Over evolutionary time, the two species may therefore be driven by selection not just to excel at controlling the resource, but also to suppress, evade, or even actively damage the other species, while simultaneously resisting the efforts of the other species to do the same. This phenomenon is called an evolutionary 'arms race', or Red Queen race (Dawkins & Krebs 1979; Holub 2001). Evolutionary arms races pervade every level of ecology, from bacterial pathogens to the largest animals: selection drives prey animals to run faster and to escape detection by predators more effectively, even as it drives predators to chase prey down more efficiently and to see, hear and smell more sharply; selection drives plants to produce tougher leaves, longer thorns and stronger poisons even as it drives herbivores to develop tougher feeding apparatus and more effective enzymes to nullify the poisons; selection drives parasitic microbes to develop ever more effective means of evading the defences enacted by their hosts and of turning those defences to their own advantage, even as it drives the host species to develop responses to stop an infection in its tracks (see Chapter **1.7** for a more detailed overview of the interactions between plants and microbial pathogens). Just as nothing in biology is said to make sense without evolution, much in ecology would not make sense without evolutionary arms races.

Evolutionary arms races may eventually come to a halt – a stable equilibrium, a 'victory' for one of the lineages (in which the other lineage either goes extinct or is permanently rendered unable to adapt to the victor's advantage), or in a perpetually repeating cycle (Dawkins & Krebs 1979). It should be noted that the latter might be considered a form of balancing selection.

Another relevant phenomenon – one that seems obvious in retrospect – is that an allele favoured by selection in one situation or environment may be at a

selective disadvantage to other extant alleles in another context. This means that as a population becomes better adapted to one set of environmental conditions, it may inadvertently sacrifice its adaptation to another set of conditions, which may not be present at that moment but may re-emerge later. This is known as an evolutionary trade-off, or the cost of adaptation, and is again a common theme across all of ecology: a population of birds trapped on an isolated island free from predators may face a selective pressure to divert resources away from developing now seldom-used flight muscles and towards competition for food or mates, but are left suddenly vulnerable when predators arrive; pathogenic bacteria bearing a gene for a metabolically expensive antibiotic-degrading enzyme may out-compete their antibiotic-susceptible relatives, but are once again at a disadvantage to susceptible strains unhindered by the metabolic cost of resistance once the antibiotic is gone. Examples of evolutionary trade-offs arising from interactions between plants and pathogens are discussed further in Chapters **1.7** and **4.1.4**.

Studies using plant species reveal trade-offs following the same principles. Populations of the invasive plant species *Ageratina adenophora* that have recently colonised areas previously uninhabited by the species, for example, have been observed to shift the allocation of resources towards growth and away from defence against herbivory than populations in already inhabited areas (Feng et al. 2009). In *A. thaliana*, the rate of production of glucosinolate – a compound involved in resistance to insect herbivory – was found to be subject to a trade-off between resistance to generalist and specialist insects. Greater production of glucosinolate was found to more effectively deter attack by generalist insects, but also to stimulate feeding and reproduction of specialist insects that had become adapted to this defence to the extent of utilising the presence of glucosinolate as a signal for promoting growth and reproduction. The evidence suggested that alleles controlling the production of glucosinolate were thus subject to balancing selection (Kroymann et al. 2003).

Evolutionary trade-offs therefore provide plausible explanations as to why variation in traits of ecological importance persists in populations. This is explored more fully in the context of disease resistance in Chapter **1.7**.

1.5.4 LOCAL ADAPTATION

Evolutionary phenomena such as arms races and trade-offs may apply uniformly across the entirety of a species' native range, though in reality this is only likely to occur in a species that is either limited to a very small range, is already adapted to a very specific habitat, or both. If the species is widely dispersed over a large area, it is highly likely to encounter many quite different habitat types. Even though a group of individuals resident in one such habitat may not be reproductively isolated from the wider population, that group is likely to evolve to become better adapted to that particular habitat type. This is local adaptation – the differentiation of a population through evolution to suit local conditions, despite the homogenising effect of gene flow.

A. thaliana is a prime candidate for a population exhibiting local adaptation: its native range encompasses a huge variety of different ecological conditions and challenges (Hoffmann 2002), and its population structure ensures that while there is sufficient gene flow to exert a homogenising influence, sub-populations are also sufficiently isolated from each other to allow adaptation to occur.

Attaining a full understanding of such a situation is a formidable undertaking, involving analyses from all of the sub-fields of life sciences discussed up to this point. Ecological work must be carried out to investigate the range inhabited by the species, identify the diversity of habitats within that range, and quantify the environmental factors likely to drive selection within those habitats. Quantitative genetic analyses must be carried out to determine differences in phenotypes and in causative genotypes between populations from different habitats. Analysis of genotypes across the range of the species must identify likely instances of selective phenomena, and must be paired with population

genetic analyses in order that effects of selection are distinguished from those of demographics.

A detailed understanding of any given instance of local adaptation is far beyond the scope of this project; and indeed – for the moment – is beyond the grasp of research on human biology. A good understanding of some specific examples of local adaptation has been attained (such as that of sickle cell anaemia as an adaptation to parasitisation in humans (Williams et al. 2005)); however, these are usually the culmination of decades of work by teams of scientists. Contributions towards some of these steps, though, are the focus of the various research chapters within this report.

1.5.5 A NOTE OF WARNING: USE AND OVERUSE OF MODEL SPECIES

While studies of model species, including *A. thaliana* have undeniably increased - and will continue to increase – our understanding of ecology, it is important to remember that model species can only take our knowledge so far. Focusing primarily on model organisms risks leaving us with a very deep but narrow field of knowledge – a superficial understanding of all life on Earth, viewed mostly through the myopic lens of a small number of species. Research into evolution and the ecology of ecosystems is a necessity to ensure that the knowledge gained in model species is applied more widely. Otherwise, how are we to know precisely how far conclusions drawn from a model species may reasonably be extended to other species? (Pigliucci 1998). Although this project has continued to focus on a model organism, it is my sincere hope that the same methods and knowledge will in the future be extended to a much wider range of species.

1.6 APPLYING 'WHOLE GENOME' THINKING IN AN ECOLOGICAL CONTEXT

1.6.1 ARABIDOPSIS THALIANA: AN IDEAL MODEL

Though work on our own species has generally led the field – indeed, many of the concepts and methods to be discussed in this report will continue to cite work on human populations that has, in one way or another, blazed the trail – most of the knowledge gained and techniques pioneered in human genetic research are equally applicable to research in other species. First amongst these other species in plant genetics has been the model plant *Arabidopsis thaliana*, or Thale cress. This species is favoured by plant scientists for its ease of growth, extensive wild range and short generation time (Meyerowitz & Pruitt 1985), as a model for molecular genetic investigation of plant development, host-pathogen interactions, and increasingly as an organism for ecological investigation of natural populations . Along with its extensive record of pre-existing genetic experimentation, its small genome – one of the smallest of any flowering plant – made it an attractive proposition to the scientists involved in the first sequencing projects; and so it was one of the first eukaryotes ever sequenced, before even our own species.

Arabidopsis thaliana is a member of the family Brassicaceae, a large and extremely diverse group that has radiated to a staggering range of habitats and ecological niches. Some wild relatives like *Capsella bursa-pastoris* (Shepherds Purse) are invasive weeds, whereas others have been domesticated as vegetable of oilseed crops such as *Brassica oleracea* (cabbage, broccoli, Brussels sprouts and cauliflower), *B. napus* (oilseed rape), *Raphanus sativus* (radish), *Brassica rapa* (turnip), and *Lepidium sativum* (cress). A better understanding of the biology of *A. thaliana* inevitably translates to a better understanding of these close relatives (Bancroft 2000).

A. thaliana is ideally suited for molecular genetic research. Its outcrossing characteristics and apparent ubiquity cause the population to form naturally

into small near-isogenic, relatively rarely intermingling groups known as 'accessions' (Mitchell-Olds 2001; Bergelson et al. 1998). Its wide native range, spanning much of the Northern hemisphere and a considerable range of latitudes, also makes it an excellent source of knowledge regarding adaptation to different climate types and habitats. Despite its propensity for self-fertilisation, the wild population retains a large amount of genetic variation, including variation at loci associated with traits of significant agronomic importance to crop species within the Brassicaceae.

Additionally, *A. thaliana* has proved a great source of knowledge in terms of genomics. The genes controlling the plant's development have been extensively mapped and studied. Much is known regarding the exact molecular mechanisms *A. thaliana* uses to detect and respond to changes in abiotic conditions such as temperature, day length and water availability in its environment. For example, GWAS experiments have revealed multiple genes associated with control of flowering time in response to day length (Ehrenreich et al. 2009).

A good deal is also known about the molecular mechanisms by which *A. thaliana* detects and reacts to challenges from biotic factors, including parasites; *A. thaliana* is known to possess both broad-spectrum defensive measures effective against a wide array of microbial species, such as those controlled by the jasmonic and salicylic acid defence pathways (Kniskern et al. 2007), and specific defenses that evolved as counters to the action of proteins secreted by pathogens in order to suppress or evade those defences. Parts of this project involving an attempt to identify instances of adaptation in *A. thaliana* will take particular note of genes of the latter type, since they are likely to be in a constant evolutionary arms race against their opposite numbers in species which parasitise *A. thaliana*. See Chapter **1.7** for an overview of our current understanding of interactions between plants and pathogens as described by the "zigzag model" (Dangl & Jones 2001; Jones & Dangl 2006),

Given this rare combination of existing knowledge and continued relevance to ongoing research, it was only a matter of time before the very latest genetic research methodology was applied to *A. thaliana*.

Figure 4 RegMap Sampling Locations from the Arabidopsis HapMap project for accessions included in the 250K SNP dataset, described in Chapter **1.6.2**. The majority of the 930 samples selected to contribute to this dataset were taken from collection sites across Western Europe, with almost all of the rest being drawn from sites in the USA or Eastern Europe. While this does not incorporate the whole of the species' wild range, this degree of sampling is sufficient to cover a broad set of habitats likely to impose different selection pressures.

1.6.2 FROM MAN TO PLANT: ONE MODEL INFORMING ANOTHER

The International HapMap project provided an impetus for the plant research community to initiate its own *A. thaliana* HapMap project in 2005. This global project, comparable in scope to its human-based equivalent, utilised a high-

throughput technique based on RNA microarrays (Borevitz et al. 2003) to genotype *A. thaliana* samples. The European and UK-wide distribution of sites at which these accessions were sampled is shown in Figure **4**. These accessions were initially genotyped using a low density set of 149 SNPs; these genotypes were then used to select a smaller subset of 916 accessions, covering as much of the population's genetic diversity as possible, which were then genotyped at a higher density of 350,000 SNPs. Many of these SNPs were rejected due to duplication or uncertainty, leaving a final dataset of genotypic variation across 216000 SNP loci, known as the 250K dataset (Kim et al. 2007).

One of the first analyses completed from this dataset was a map of linkage disequilibrium across the *Arabidopsis* genome. Despite the low outcrossing rate of the species, Kim *et al.* (2007) found that linkage disequilibrium between loci breaks down when loci are separated by, on average, 10kb.

1.6.3 A. THALIANA AS A MEANS OF REVEALING ECOLOGICALLY IMPORTANT VARIATION

As with human genomics, genome-wide association studies have been performed for many ecologically and agronomically significant traits in *A. thaliana*. A prominent example is that of flowering time. Since *A. thaliana* inhabits such a surprisingly vast range of latitudes – from Scandinavia to the sub-tropics – it must be able to adapt to a wide variety of climatic temperature ranges and day lengths. In fact, *A. thaliana* is known to have developed different strategies for the combinations of these and other climatic variables it faces across its latitude range, and control of flowering time is known to play a key role in this adaptation (Michaels et al. 2003). Across most of its range, *A. thaliana* typically lives as a winter annual, producing a single generation each year. Growth begins with autumnal germination and continues the winter, terminating with flowering and seeding in the spring and seed dormancy over the summer. Towards the more northerly and colder extremes of its range, however, *A. thaliana* possesses alleles associated with a summer annual lifecycle (Michaels et al. 2003; Alonso-Blanco & Koornneef 2000). The genetic

basis of this alternative life cycle is known to involve variation in genes that control vernalisation and flowering time, as described by Michaels *et al.* (2003) Consequently, genes associated with this trait, and other genes associated with flowering time, are likely candidates for local adaptation. Analyses described in this project therefore sought signatures of selection acting upon these genes (see Chapter **4.3.3**).

1.7 HOST-PARASITE INTERACTIONS AS A MODEL FOR LOCAL ADAPTATION

1.7.1 THE ZIGZAG MODEL OF PLANT-PATHOGEN INTERACTIONS

Interactions between plants and their microbial parasites provide us with an excellent means of examining two aspects of evolutionary theory with considerable importance to ecology: evolutionary arms races between host and parasite, and the process of adaptation to changing external factors. Depending upon the species under investigation, plants may be grown in large numbers using relatively simple, inexpensive facilities; pathogens may be applied readily and consistently to the hosts; and the interactions between the pathogens and hosts may be easily observed, since the hosts are not mobile. Although informative from a purely theoretical perspective, research into interactions between plant hosts and pathogens also leads to practical benefits in terms of the protection of food crops against destructive epidemics – which, indeed, is often the primary motivation for following a line of research using particular hosts or pathogens.

For decades, our understanding of the genetics of plant-pathogen interactions has been shaped by the highly successful gene-for-gene model (Flor 1971). Flor reported that the outcome of interactions between host and pathogen species was determined by the presence or absence of paired genes; *R* (or resistance) genes in the host, and *Avr* (or avirulence) genes in the pathogen. The presence of *Avr* genes was demonstrated to enable a pathogen to cause infection by the cloning of *Avr* genes into avirulent pathovars – notably in the case of *Pseudomoas syringae* (Kobayashi et al. 1989).

The modern understanding of interactions between plants and pathogens is represented in the "zigzag model" (Dangl & Jones 2001; Jones & Dangl 2006). Under this model, interactions between plants and pathogens are represented as a series of back-and-forth adaptations representing an evolutionary arms race. These adaptations occur in 4 sequential phases (see Figure **5**). This model

allows us to understand the co-evolution of plants and pathogens from the perspective of both parties (Dodds & Rathjen 2010).

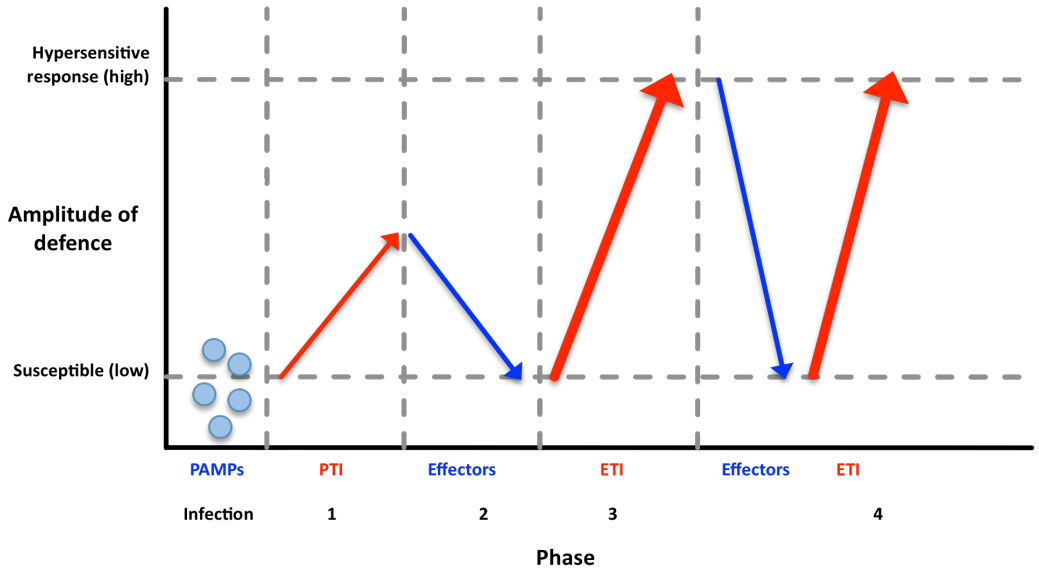


Figure 5 The Zigzag model of plant-pathogen interactions The zigzag model describes the progression of an evolutionary arms race between plants and pathogens, in which pathogens are forced to evolve new effectors in order to evade or suppress host immune responses, which in turn forces plants to develop new effector-triggered receptors and strategies. The presence of pathogens is first detected when pathogen-associated molecular patterns (PAMPs) are recognised by the host's pattern recognition receptor (PRR) proteins, and defence signalling pathways are activated. This is termed PAMP-triggered immunity (PTI) (**Phase 1**). Pathogens may evolve to suppress this response by producing effector proteins, which inhibit the pattern recognition receptors or defence pathways, causing effector-triggered susceptibility (ETS) in the host (**Phase 2**). Hosts evolve to produce R proteins, which detect the suppression of PTI, and are therefore termed effector-triggered immunity (ETI). R proteins directly induce defence responses and activate effector-triggered immunity, including programmed cell death in areas surrounding sites of infection (the hypersensitive response) (**Phase 3**). Pathogens then undergo selection favouring mutations in the effectors that suppress PTI but do not trigger ETI, and may evolve a second effector to suppress ETI. Hosts are likewise under opposing selection pressures to detect suppression of immunity via ETI (**Phase 4**).

Following infection, the host presence of pathogens is first detected when pathogen-associated molecular patterns (PAMPs) are recognised by the host's pattern recognition receptor (PRR) proteins, and defence signaling pathways, are activated. This is termed PAMP-triggered immunity (PTI). PRR kinases are typically membrane-bound, in order that pathogens invading the plant's intracellular spaces can be recognised (Monaghan & Zipfel 2012). Upon recognising molecular motifs consistent with the presence of pathogens, PRR proteins trigger numerous defence-related signaling pathways, including the salicylic and jasmonic acid pathways (Loake & Grant 2007).

PAMPs are generally motifs that pathogen species cannot readily alter, such as bacterial flagellin, which is recognised by the FLS2 receptor (Gómez-Gómez & Boller 2000). The evolutionary opportunity for pathogen species to avoid detection through alteration of PAMP motifs is usually constrained by the necessary functions of those proteins or molecules – for example, flagellar motility is crucial for high infectivity of *P. syringae* (Panopoulos 1974), so genetic variation sufficient to guarantee evasion of recognition by PRRs would also impact the ability of the pathogen to cause infection. While some variation in PAMPs and corresponding PRRs has been observed (Gómez-Gómez et al. 1999; Felix et al. 1999), the degree of constraint on PAMP variation frequently necessitates the evolution of other means of evading host defences.

Pathogens may instead evolve to suppress PAMP-triggered immunity by producing effector proteins (coded by *Avr* genes), which inhibit the PRR kinases or defence pathways, causing effector-triggered susceptibility (ETS) in the host. Unlike PAMPs, which are frequently released from sites of infection by diffusion (for example, fragments of flagellin released from dead bacteria diffuse away from the infection site) (Haefele & Lindow 1987), effectors are commonly transported or otherwise actively deposited directly into host cells.

For example, *P. syringae* and other pathogenic bacteria use a protein complex – the Type III protein secretion system – to transfer effector proteins to the inside of host cells (Büttner & He 2009). Once inserted into the host cell, the introduced effectors interfere with the host's signaling pathways, preventing the induction of PTI. Amongst the proteins inhibited in order to prevent the triggering of PTI is *RIN4*, which is inhibited by the effectors *AvrB*, *AvrRpt2* and *AvrRpm1* (Kim et al. 2005).

When pathogens become capable of inducing ETS, hosts are subjected to a selection pressure favouring individuals capable of counteracting the effector. Hosts thus evolve *R* genes, producing R proteins (or, indirectly, other molecules involved in defence responses or signaling), which detect the suppression of

PTI by effectors, and are therefore termed effector-triggered immunity (ETI). Most R proteins possess leucine-rich repeat (LRR) domains, and many also have nucleotide-binding domains (Dangl & McDowell 2006). Unlike PRR proteins, R proteins tend not to have transmembrane domains, indicating that they do not interact with their targets at the cell membrane, but within the cytosol (Dangl & Jones 2001). Well-studied examples of R proteins are those encoded by genes *RPM1* and *RPS2*, which recognise the inhibitory actions applied by effectors to the protein produced by gene *RIN4* (Kim et al. 2005).

Upon detecting its corresponding effector, an R protein triggers signalling pathways causing the plant's defence mechanisms to become more active. R proteins generally also induce the 'hypersensitive response', in which infection is halted through the apoptosis, or programmed cell death, of cells surrounding the immediate site of infection. (Nimchuk et al. 2003; Mur et al. 2008).

Pathogens then undergo selection, in which variant effectors that suppress PTI but do not trigger ETI are favoured. A second effector to suppress ETI may also evolve. Hosts are likewise under opposing selection pressures to detect suppression of immunity via ETI. As is apparent from the example of the multiple effectors and multiple *R* genes acting upon *RIN4*, this is likely to result in the co-evolution of complex interaction networks, in which many genes from both host and pathogen contribute to an overall phenotype. Moreover, in reality it is possible for plants to come under attack from more than one race or species of pathogen simultaneously; in such a case, the immunity suppression measures implemented by one pathogen, and the defences mounted by the against it, are likely to affect the infective ability of the second pathogen (see Chapter **4.1.4**) While this realisation moves the field away from the simplicity of the gene-for-gene model which produced the original insights into the genetics of plant-pathogen interactions, it also opens up the possibility of new insights from the application of quantitative genetic thinking, as discussed in Chapter **1.4**.

1.7.2 EVOLUTIONARY ARMS RACES BETWEEN PATHOGENS AND HOSTS

It is the R protein-mediated defence responses that are of greatest interest to this project, and to ecological research as a whole. On the evolutionarily short-term scale of adaptation to the immediate environment, the scope for the evolution of new chemical defences is relatively limited, and as previously discussed in Chapter **1.7.1**, variation in PAMP motifs (and consequently also in the PRR proteins responsible for recognising them) tends to be more limited than variation in effectors. Instead, evolution has been shown to act intensely on variation within *R* genes (Holub 2001). *R* genes are known to be subject both to evolutionary arms races with their *Avr* gene counterparts, and to several other evolutionary trade-offs. It follows that *R* genes would be expected to exhibit signatures of recent selection.

R proteins have been found to have two general modes of action: either binding effectors directly, or recognising alterations to signaling pathways induced by effectors (as per the 'guard hypothesis' (de Meaux & Mitchell-Olds 2003; Dangl & McDowell 2006). The two modes of action have differing evolutionary consequences for the host and the pathogen. In the former case, since *Avr* genes are under selection pressures to evade or inhibit detection, the *R* gene must also be under diversifying selection, leading to the emergence of novel variation in R proteins in order that novel variation in effectors may be detected. This may lead to the establishment of balancing selection through a form of frequency-dependent selection; rare *Avr* alleles would give a pathogen an advantage, allowing it to overcome ETI due to the fact that few hosts would possess R proteins capable of recognising the *Avr* protein. As this *Avr* allele rises in frequency in the pathogen population, *R* genes producing proteins capable of recognising it would be at a greater selective advantage, and would therefore also increase in frequency in the host population. This would in turn nullify the selective advantage of the *Avr* allele, causing its frequency in the pathogen population to decrease – followed, once more, by the frequency of the *R* gene allele in the host population as its target becomes less common (Thrall &

Burdon 2003; Tian et al. 2003). Since the presence of allelic variation in the host population is necessary for this situation, it follows that *R* genes coding proteins following this line of action are likely to be found at the loci most frequently undergoing recombination in meiosis.

On the other hand, in cases in which R proteins monitor other host proteins for interference caused by effectors, the opposite is likely to be the case. Since the targets of effectors are unlikely to be under strong diversifying selection, *R* genes are likely to be under a degree of stabilising selection – selecting against alleles producing R proteins incapable of recognising the inhibition of defence pathways by effectors. *R* genes producing proteins following this mode of action could therefore generally be expected to demonstrate selection by sweeps rather than balancing selection, since (less functional) variant alleles of this *R* gene are not advantageous at any point – unless there is a selective advantage in failing to initiate a defence response.

In terms of evolutionary trade-offs, there is a selective benefit to the host in only activating its defence mechanisms when genuinely under attack by a pathogen; defence responses are metabolically expensive (Heil et al. 2000; Heil 2002; Tian et al. 2003) and may be damaging towards both the plant (due to the tissue death inherent to the hypersensitive response) and any beneficial symbionts resident on its surface or within its tissues (Kniskern et al. 2007).

Therefore, there is a selective pressure against any propensity for *R* genes to suffer 'false alarms'. Alleles less susceptible to false alarms, however, make the trade-off of being less likely to detect variants in the pathogen avr genes, and therefore potentially less able to activate the defence pathways when the plant is under attack. In terms of selection, this is likely to mean that variation in this characteristic is maintained in the gene pool of the wild population due to the pathogen challenging the host only intermittently, rather than constantly; there are times when the less sensitive *R* gene allele is favoured due to its reduced

costs of false alarms. Real-world examples of precisely this evolutionary trade-off have been encountered in *A. thaliana* (Todesco et al. 2010).

Consequences of this trade-off extend still further. While the hypersensitive response is generally effective at preventing the progression of infection by obligate biotrophs or hemi-biotrophs such as *Pseudomonas syringae*, cell death actually assists infection by necrotrophs – pathogens that kill host tissues before feeding. Govrin and Levine (Govrin & Levine 2000) demonstrated that the necrotrophic fungal pathogen *Botrytis cinera* induced the hypersensitive response as part of its infection strategy; and also that when the hypersensitive response was induced in order to successfully defend against challenges from biotrophs (here, *P. syringae*), *B. cinera* was able to exploit the necrotic tissue produced by the hypersensitive response in order to bring about a highly damaging degree of infection. Adaptations resulting in a successful defence against one pathogen may therefore represent a trade-off in fitness against attack by other pathogens that have become adapted to overcoming that defence.

1.7.3 ECOLOGICAL ASPECTS OF EVOLUTIONARY ARMS RACES

Arms races may also have ecological aspects. Much of the ecological theory in the field involves the case of an invasive species successfully colonising a new habitat, in which the predators and pathogens of its home range (that the invader is now said to have 'escaped' from) are at least temporarily absent (Blossey & Notzold 1995). In this new environment selection is likely to favour alleles causing the organism to devote a greater proportion of its resources toward growth and reproduction, and away from defence, that would have been selected against in the organism's home range due to the prevalence of disease-causing parasites. Experiments using artificial clines have shown that this 'fitness cost' is a real phenomenon (Roux et al. 2006). Under these circumstances, evolutionary arms races arise between individuals within a

species competing to dedicate greater resources toward growth, and also between native species inhabiting the same or similar ecological niches.

In the real world, both hosts and pathogens are also likely to face yet another set of trade-offs: pathogens in the wild frequently infect multiple hosts, and hosts will often experience attack from multiple pathogen species – often simultaneously (Kniskern et al. 2007; Barrett et al. 2009). This could be predicted to lead to a trade-off between host range and likelihood of success on any individual host species on the part of the pathogen, and vice versa (a trade-off between adequate defence against multiple pathogens and superior defence against just one) on the part of the host.

However, the infective abilities of a pathogen are determined as much by abiotic conditions of the environment as by the genetic properties of the host (Parker & Gilbert 2004). Successful infection of a host requires three equally important conditions, forming the 'disease triangle': susceptible host, virulent pathogen and suitable environmental conditions (Scholthof 2007), since pathogens are usually only capable of successfully infecting a host within a specific range of those conditions.

Together with the ecological aspects of arms races mentioned above, this means that resistance to disease is an ecological matter as much as an evolutionary one. Exactly how a given range of environmental conditions might impact upon this trade-off in a real-world system remains an ongoing topic of research in the field of ecological genetics – one with a pressing requirement for sound understanding as the effects of global climate change begin to become noticeable, potentially altering the worldwide distribution of plant pathogens in an economically significant and otherwise unpredictable manner. Knowing which genes (and which alleles of those genes) are responsible for resistance to particular disease-causing pathogens, then, may provide tangible benefits to applied research such as crop breeding programs and agricultural epidemiology (Gilligan 2008). Knowledge of how evolution between hosts and

pathogens proceeds in given ecological contexts should also guide the agricultural policies of governments in order to best manage the risk pathogens pose to a stable food supply.

In addition to those practical benefits, knowledge of ecological aspects of selection helps to provide answers to the more theoretical open questions in ecological genetics. Plant-pathogen interactions provide a powerful insight into the progression of evolution across differing ecological conditions as the host species becomes adapted to its local conditions. Does evolution of disease resistance primarily proceed through a continual, endless arms race, in which new resistance alleles are swept rapidly to fixation within a habitat? Or do resistance alleles endure through time and across habitats in the stable equilibrium of balancing selection? By attempting to provide answers to this question, this project aims to explain why variation exists within the genotypes governing a key trait to any organism's survival and success.

1.8 THE CASE FOR THIS PROJECT

The field of ecology must be able to expand its remit into a full understanding of how the variation between genotypes translates to the success of populations in the complex and variable combinations of factors comprising real-world environments. The work presented in this project aimed to contribute towards this level of comprehension. Due to the historical direction of the field, work has focused largely on using 'model species' as means of testing proposed hypotheses. This has clearly proved effective in many respects, since it has revealed a number of key ecological principles (see Chapter **1.5**), though it leaves us relatively poorly equipped to understand the simultaneous effects of genetic variation across individual phenotype, fitness to an immediate habitat, and fitness to the complex, constantly changing web of interactions constituting an ecosystem as a complete system. In studying the interactions between a host and a parasite species in the context of a wild population – a relationship open to several levels of evolutionary trade-offs, as discussed in Chapter **1.7** – this project aimed to provide a means of investigating and understanding these complex interactions.

The modern availability of high-quality genotype data and analytical tools made it possible to conduct research into not only the demographic past of a wild species, but also the ongoing selective challenges the species faces from abiotic and biotic factors in its environment. However, the relatively recent advent of these advances in data and methods means many questions remain unanswered or incompletely answered. In particular, the exact details of the population genetics and adaptive challenges faced by the *A. thaliana* population in the United Kingdom remain an open question. The detailed nature of plant-pathogen interaction research around the UK opens up a fascinating window into the evolutionary world of a wild population, and methods developed in the course of investigating this specific circumstance are likely to also find applications in the many other areas in which our understanding of the processes and drivers of adaptation are as yet incomplete.

In its initial stages, this project marks an opportunity to re-examine phenomena reported by previous work in the light of newer, more dense data with the goal of testing prior predictions. The existence of the 250K SNP dataset granted an opportunity to apply the most thorough test yet carried out for the existence of the still contentious 'isolation by distance' population structure in *A. thaliana*. If, as Platt et al. (Platt et al. 2010) reported, this type of population structure was apparent, the marker density of the 250K dataset allowed an unprecedented quantification of the degree to which that structure exists across various parts of the species' range. If, however, the general structure of the population follows an alternative pattern, that structure would also have been readily identifiable from the 250K dataset.

Research presented in this project also fills in gaps in published knowledge relating to the distribution and historical migration of *A. thaliana* genotypes. Work by Horton *et al.* (Horton et al. 2012) showed that the UK *A. thaliana* population either originated from several sources or secondarily founded other populations, but did not elaborate upon these findings. This analysis aimed to repeat the observation, and then to expand upon Horton *et al.*'s (2012) analysis by examining the possibility that genotypes arising from different sources are, or are becoming, adapted to specific habitats.

These analyses facilitated the generation of a simple model of seed and pollen dispersal capable of reproducing the population structure observed from the 250K dataset. Just as a comparison between observations and the predictions of Hardy-Weinberg equilibrium inform us of the causes of observed phenomena, this research aimed to distinguish genotypic signatures of selection from the background of selectively neutral genotypic variation by comparing haplotypes observed in the 250K data with expectations generated from an application of this model, and in so doing, provided a genome-wide overview of selection acting upon a wild population. It was then possible to examine specific cases of

selection acting upon loci known to be of significance to phenotypic variation in defence against pathogens and in adaptation to specific habitat conditions.

The methods and results presented in this project aimed to demonstrate that a synthesis of ecological, geographic and genotypic data can lead to a powerful means of producing knowledge, and can guide future research in productive and novel directions.

1.9 CAVEATS AND WARNINGS

Over the course of this project, many of the major conclusions are drawn from analyses based around a demographic model representing a selectively neutral population, or from statistical deviations of observed genotypes from expected values derived from that model. Due to the difference in number of individuals between the wild population extant in the UK and the number of individuals sampled for the 250K dataset, it was necessary to adjust several of the parameters of the model, in order that the model conformed to the observed characteristics of the wild population. (See Chapters **3.3.2** and **3.3.3** for details, and Chapter **3.4.2** for discussion).

Altering parameters in this manner risks causing the model to fail to conform to the nature of the real population. Verification of the results obtained using this model was therefore undertaken where appropriate, in order to ensure that - as far as possible – conclusions drawn from it remained consistent with observations of the wild population after the parameter scaling was applied.

1.10 WHOLE PROJECT PLAN OF ATTACK

Inherent widespread admixture and population structure in *A. thaliana* have, on occasion, been cited as reasons against utilising the species in the context of ecological investigation. By investigating a model of population structure in this species that is then used as a baseline for testing of ecological and evolutionary hypotheses, I intended to demonstrate with this project that *A. thaliana* is also a useful and powerful model for ecological research.

Over the course of this project, I have aimed to develop new methods of identifying loci under selection, and of looking into the migratory past of a wild species; I have also sought to develop software to assist in the application of those methods – tools which have been, as far as possible, created with a broader range of applications than the immediate scope of this project in mind.

The aims of this project were as follows:
A1: Resolve in greater detail the geographic sources of the various genotypes present in the UK *Arabidopsis thaliana* population. (Part **A**, Figure **6**)
A2: Construct an ecological model from which the amount of time passed since the UK population first became established can be estimated. (Part **B**, Figure **6**)
A3: Apply that ecological model to the identification of genomic signatures of local adaptation to the particular habitats in which *Arabidopsis thaliana* is found in the UK. (Part **C**, Figure **6**)

Consequently, the hypotheses this project aimed to test were set up as follows:
H1: High-density haplotype analysis shows that the structure of genetic variation in *A. thaliana* populations follows an isolation by distance model. (Part **A**, Figure **6**; Aim **A1**)
H2: The UK population arose from a single source of founders on the European mainland. (Part **A**, Figure **6**; Aim **A1**)
H3: The UK population has been established for roughly 1000 years. (Part **B**, Figure **6**; Aim **A2**)
H4: Local groups of *A. thaliana* within a particular habitat possess alleles at a

frequency significantly different from that expected under a selectively neutral model, indicating that selection is acting upon those alleles. (Part **C**, Figure **6**; Aim **A3**)

H5: Signatures of selection indicating that selection is acting upon alleles associated with disease resistance are shared consistently across populations in different habitat types. (Part **C**, Figure **6**; Aim **A3**)

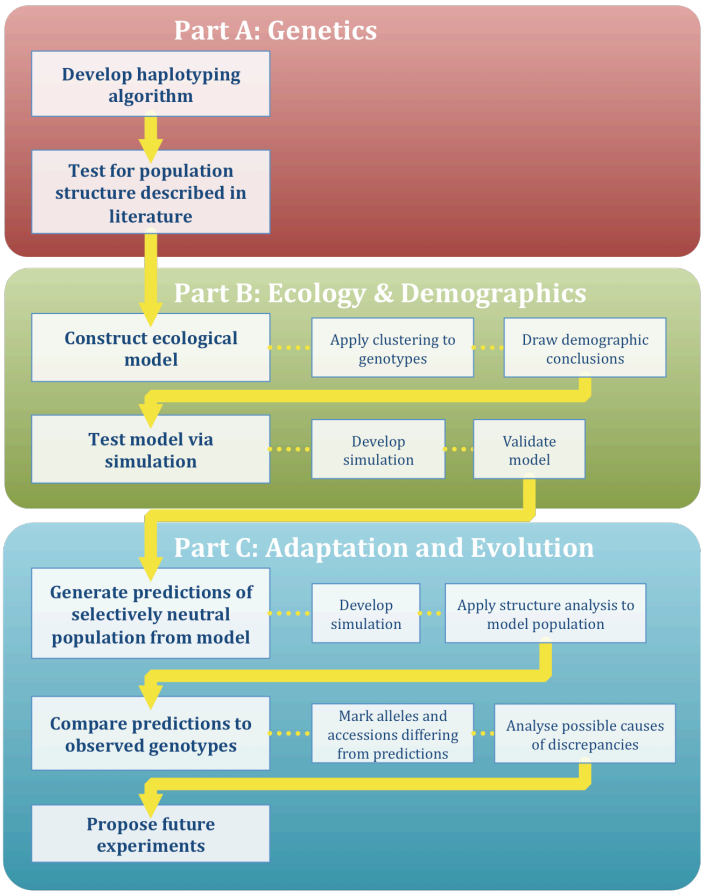


Figure 6 Whole Project Overview. Flow chart showing how the component parts of this project fit together to form an analysis pipeline. Boxes branching to the right (smaller text) represent tasks required for the fulfilment of the project's sequential goals (larger text), which may be read in downward order for the quickest overview. The first part of the project (Part A) concerns the identification of conserved sets of alleles (haplotypes) from high-density genotypic data taken from a wild population. Standard population genetic analyses are used to reveal structure within the extant population. The second part (Part B) involves the construction of a model of individual dispersal and reproduction capable of providing an approximate explanation of the observed population structure. In the same way that differences from the predictions of Hardy-Weinberg equilibrium are informative, the third part of the project (Part C) concerns departures from this model. Significant departures from the model may represent instances of a sub-population becoming adapted to a specific habitat type. Potential causes for these departures from the model are suggested, and further experiments are proposed in order to investigate these possibilities more extensively. Parts A, B and C are the subject of chapters 2, 3 and 4 respectively.

In order to fulfil these aims and test these hypotheses, I have created a set of software tools to analyse and manipulate high-density polymorphism data all the way from raw SNP data to being able to provide meaningful answers to the questions set out above.

A flow chart depicting the overall process of the project is shown in Figure **6**. Haplotypes must first be detected from the SNP data; the method I used is described in Chapter **2**. Since the geographic locations of the accessions possessing each haplotype are known, it is possible to examine population structure once haplotypes have been identified (as discussed at the end of Chapter **2**), and also to putatively identify the sources of genetic variation within an area (Chapter **3**).

Once the population structure is known, it becomes possible to undertake more complex analyses. Since isolation by distance-type population structure is a phenomenon that emerges as the population becomes more established following its founding, the degree to which the structure has emerged may give a clue to the amount of time that has passed since that founding event. A simulation-type approach was used to create a model of the emergence of this population structure, and thereby to estimate the number of generations the current UK population may have been resident.

Similarly, a sound model of population structure allows for an analysis designed to identify recent and ongoing local adaptation (Chapter **4**). In the final analysis of the project, the set of haplotypes found in the UK population was compared against a set of haplotypes generated from a model population subjected to complete selective neutrality. This analysis was, essentially, a GWAS; looking for alleles associated with non-neutrality. Again, since the geographic spread of haplotypes was known, spatial data was also used to add power to the analysis.

CHAPTER 2: IDENTIFICATION OF HAPLOTYPE BLOCKS

2.1 INTRODUCTION

2.1.1 PLAN OF ATTACK

Work described in this chapter was aimed at testing hypothesis **H1**: High-density haplotype analysis shows that *Arabidopsis thaliana*'s population structure follows the isolation by distance model, specifically in the UK compared to other regions of Europe.

The goals of this chapter were two-fold. Firstly, to develop a method for reliable and rapid means of gathering haplotype data from genomic data, and then to apply this method to the available *A. thaliana* 250K dataset (Kim et al. 2007). This itself served a dual purpose, both to convert the large dataset into a format more easily managed and interrogated (a characteristic of ever-increasing importance as genomic datasets continue to rapidly rise in size and complexity), and to facilitate further investigations into local adaptation in subsequent chapters. Applicability of the approach to other datasets and situations was considered a desirable outcome of the final approach.

The second goal of this chapter was to investigate characteristics of the wild *A. thaliana* population including population structure, upon which later analyses would also be based. To this end, the haplotype data recovered through the fulfilment of the first aim was used to re-evaluate the applicability of the isolation by distance model proposed by Platt (Platt et al. 2010) to explain patterns of genetic variation amongst natural populations of the wild species.

2.1.2 DEFINITION AND CHARACTERISTICS OF A HAPLOTYPE BLOCK

Throughout this project, the term "haplotype" will be used to refer to a single pattern of alleles potentially shared across a number of individuals, as

described further below. The term "haplotype block" will be used to refer to a set of *A. thaliana* accessions possessing a particular haplotype.

A haplotype is a set of alleles that span a number of loci in a single segment of chromosome that are very tightly bound together by linkage, and therefore are found consistently together in a specific pattern. This linkage is essentially an artefact of the finite amount of recombination that occurs during meiosis; were crossovers to form between every single possible locus along the length of a pair of homologous chromosomes, haplotypes could never exist.

Over time, the allelic combination of a haplotype is gradually broken apart by recombination during meiosis. While a given haplotype may originally arise and spread through a population as a combination of alleles spread over an entire chromosome, recombination will soon cause other combinations of alleles to be interspersed along its length. Since a haplotype will usually be shared amongst a number of individuals within a population, the original whole-chromosome haplotype can be expected to be broken apart into smaller fragments; some individuals may retain a shared set of alleles at one end of the chromosome, for example, while a second group of individuals may retain another set of alleles near the other end of the chromosome. Given complete and random reproductive mixture of the population (a state essentially of Hardy-Weinberg equilibrium), it is a statistical certainty that a haplotype will become shorter and eventually be completely broken down by the action of recombination. Random mutations also play a role in breaking up haplotypes, though to a much lesser degree than meiotic recombination.

An important consequence of this recombinatory breakdown is that individual organisms may not share exactly the same parts of a given haplotype. Under conditions of Hardy-Weinberg equilibrium, a random yet ultimately predictable process (representable mathematically as a trend of exponential decay of mean haplotype length) determines the proportion of a haplotype any given individual possesses. The unpredictable nature of meiotic recombination means

that the exact fraction of the haplotype an individual receives from its parents cannot be known *a priori*, yet the fraction of the original whole-chromosome haplotype remaining in the population will be, on the whole, predictable over time. If neutrality is assumed, then, it becomes possible to estimate the age of a haplotype based on its average length within the population and the size of that population (Kimura & Ota 1973). The consequences of this breakdown phenomenon in terms of selection will be explored more fully in Chapter 4.

2.1.3 EXISTING METHODS AND ALGORITHMS

While the computing power available to geneticists has grown at least as rapidly as the amount of data available to them, more advanced applications of genetic science on a whole-genome scale often remain computationally demanding.

Processing time for comparisons between large numbers of samples across hundreds of thousands to millions of loci is usually subject to polynomial time – the time to complete all comparisons scales exponentially as the number of samples scales linearly. Despite steady advances in computational speed and capability, consideration of polynomial processing time is rarely far from the mind of a programmer creating tools for genomic analysis. Consequently, many genomic analysis tools are developed with lower processing time – as well, of course, as superior precision, accuracy and robustness to error – as one of their stated goals or advantages. This project is no exception; in this section, I have aimed to create a method of finding haplotype structure from high-density polymorphism data that is reliable, simple to use, and *fast*.

The method developed in this chapter drew inspiration from a paper by Zahiri et al. (Zahiri et al. 2010). They proposed a theoretical principle for identifying haplotypes within a large volume of genomic data, and provided a mathematically sound method based around two functions described as IsBlock and NeighbourBlock.

IsBlock simply checks whether a section of the genome – for example a multiple sequence alignment of part of the genomes of several samples from within a population – can reasonably be considered as a haplotype. The exact criteria by which IsBlock determines this are left open (so may in practice include a simple measurement of genetic diversity such as number of pairwise differences, more advanced distance measures based on variant substitution rates, or measurements of linkage disequilibrium over the whole set of loci). What matters most is that it is possible to set a clear, even if arbitrary, threshold of genomic similarity that can be applied consistently. An initial scan of the data using IsBlock identifies clusters of variation spanning small regions of the genome; a sliding window approach may be taken, in which each subset of sample variation within a small range of the loci covered in the data is initially tested with IsBlock.

The NeighbourBlock function seeks to extend existing clusters identified by IsBlock into longer haplotype blocks, by combining cluster groups at adjacent loci. The combined genomic data are then passed once more to the IsBlock function. Should IsBlock report that the new cluster still qualifies as a haplotype, NeighbourBlock joins the two cluster groups together as appropriate. Extension by adding clusters thus continues until there remain no further clusters that, when integrated into the existing cluster, still register as a valid haplotype. Making multiple passes over the cluster group data to ensure all possible clusters are joined is recommended. A mathematical proof that the clusters derived from the shared sequences logically confirm or dissuade the consideration of combined cluster groups as a single block is presented in the paper.

Utilising this method, Zahiri *et al*. (2010) report finding haplotypes from genotype data both more accurately and more quickly than with other comparable methods.

2.2 MATERIALS & METHODS

2.2.1 IDENTIFICATION OF HAPLOTYPES

Almost all of the tools developed over the course of this project were written in Perl. This high-level programming language is designed for general-purpose application. Its main strengths lie in its built-in tools for the manipulation of text strings (for example, sequences), including regular expression-based pattern matching, character substitution and string division. Perl programming is memory-intensive yet its relatively intuitive features such as lists, automatic memory management and hashes make it easy to learn; its extensive repository of publicly licensed modules and add-ons (many of which are expressly designed with scientific applications in mind) make it accessible and adaptable. All listed Perl modules are available from the Comprehensive Perl Archive Network (CPAN – the central public repository of Perl software (http://www.cpan.org/)), and are typically designated by their general class of function followed by the specific task (or set of tasks) for which they are designed, separated by a double colon mark.

The 250K *A. thaliana* HapMap dataset was divided into 'windows' of arbitrary length (typically 50 or 100 SNPs) using a perl script. Care was taken to ensure that no window contained SNPs from two different chromosomes. The level of similarity between SNP-sequence alignments within each window was measured using the bioinformatic analysis program dnadist (Felsenstein 2002). Hierarchical clustering on the distance matrix produced by dnadist was carried out using the UPGMA method, using the module Algorithm::Cluster (Hoon et al. 2004). An arbitrary cut-off was imposed on this hierarchical clustering using the module Algorithm::Cluster::Thresh. Clusters from individual windows were joined using the Kuhn-Munkres algorithm (Munkres 1957), implemented in perl through the module Algorithm::Munkres. Groups of clusters that were considered at this point as haplotypes were stored in a newly developed format, which recorded a list of all accessions possessing each haplotype, and also the contiguous windows within which each accession actually possessed

the haplotype (hence referred to as a 'run'). Groups of clusters were split apart again, where appropriate, by first applying the Shapiro-Wilk normality test (SHAPIRO & WILK 1965) (taken from the module Statistics::Normality) to the midpoints of all the runs associated with each haplotype and then, if the normality test returned a result indicating that a normal distribution of run-centres is unlikely, by applying K-means clustering (MacQueen 1967) (implemented in the module Algorithm::Kmeans) to the set of run-centres.

2.2.2 ANALYSIS OF POPULATION STRUCTURE

In addition to perl, much of the analysis carried out in this project utilised data-handling, statistical and graphing functions of the R high-level programming language. As with perl, its abundance of science-oriented modules available from its Comprehensive R Archive Network (http://cran.r-project.org/) public repository, its relative ease of use and its enormous versatility render it a popular tool in the scientific community.

Charts showing results were created using either R or perl scripts utilising the module Image::Magick. Population structure was estimated by measuring the genetic dissimilarity, in terms of the sharing of haplotypes, between accessions at a range of distances. Distances between accession collection sites were calculated from map coordinate data associated with each accession using the Haversine method (Robusto 1957) with the module GIS::Distance. Trends from this data were analysed using a linear regression implemented in module Statistics::LineFit, which also calculated R^2 values.

All processing for this analysis was carried out on a 2009-built dual-core MacBook Pro with 8GB RAM.

2.3 RESULTS

2.3.1 A FAST, CLUSTER-BASED APPROACH FOR THE IDENTIFICATION OF HAPLOTYPES

Much of the work in this chapter was focused around the development of a tool aimed at rapidly identifying haplotypes from high-density SNP data collected from a large number of samples. This tool went through several iterations and refinements before a final version combining many of the best features developed in previous versions was settled upon. That development process, and the reasons behind the decisions made over the course of development, will be addressed here. These decisions had a substantial impact on the rest of the conclusions drawn from this project (as briefly described in Chapter **1.9**), since the data describing haplotypes within the population was used as input to the subsequent analyses described in Chapters **3** and **4**.

A fundamental issue in any attempt to identify haplotypes must first be described. The existence of haplotypes implies that neutral variation, at least in general, predicts the presence of other alleles at nearby loci (Charlesworth et al. 2003). This is, however, a trend rather than an absolute rule. Situations in which variation might deviate from the predictive trend are conceivable, though the presence of recent genetic variation at a locus due to mutation are usually more plausible than any explanation rooted in meiotic recombination, since crossovers only rarely ooccur in extremely close proximity to each other (Griffiths et al. 1999). In order to account for these possibilities and reveal the underlying trend of mutually predictive variation, then, some flexibility must be allowed in any method aiming to detect haplotypes. The method developed for this chapter allowed for this flexibility by clustering small groups of alleles spaced near to each other along the genome, rather than examining individual loci independently. The rest of the surrounding sequence, which will (in the absence of other simultaneous new variation) continue to hold the haplotype if one exists, will still enable the correct clustering of that part of the sequence with other samples carrying the same haplotype, despite any errors or small-

scale variation such as recent mutation events. Figure **7** shows how the solution ultimately chosen for this project accommodates this possibility.

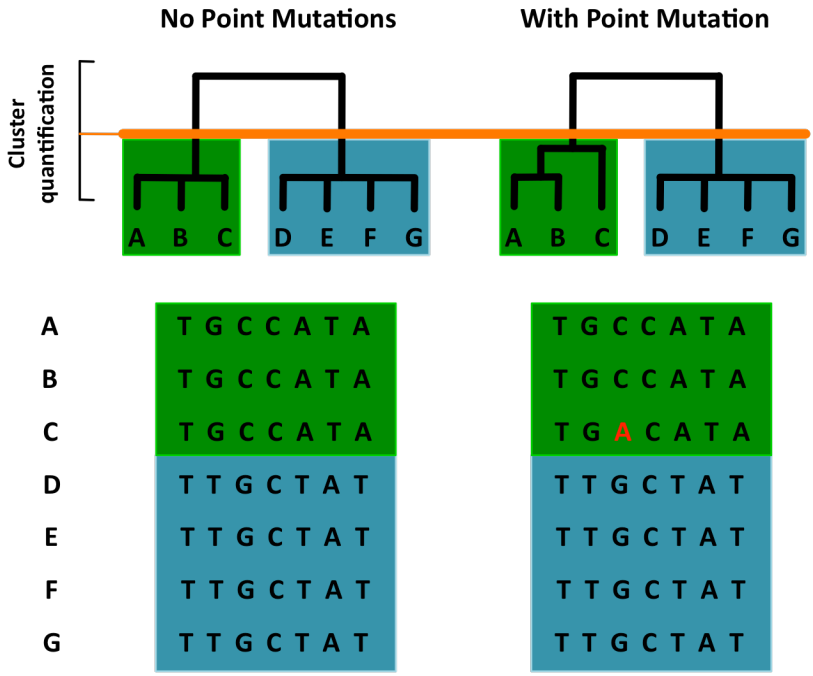


Figure 7 Recognition of haplotypes despite recent point mutations Changes in allelic composition of a set of loci may be attributable to recent point mutations rather than the meiotic recombination responsible for the degradation of haplotypes. A means of distinguishing recent mutations from underlying haplotype structure is implemented as a side-effect of the hierarchical clustering technique used to identify the genotype clusters (see Chapter **2.3.1**). This figure shows how applying a threshold to a hierarchical clustering allows the identification of haplotypes despite the existence of non-recombinatory variation (i.e., point mutations – **RED text**) in a SNP sequence. The threshold applied to the hierarchical clustering (**ORANGE line**) allows for a small amount of variation between SNP sequences, resulting in the haplotype groupings being recognised despite point mutations.

It is possible that any new mutation arising and propagating within an existing haplotype could form a completely new one, encompassing the whole of that chromosome, and that the analysis should therefore use this definition. However, this strict definition is not practical when using the 250K dataset, because the density of SNPs was developed to select ca. 1100 samples from more than 6000 available accessions (see (Platt et al. 2010; Horton et al. 2012)) in order to reflect the genetic diversity of the entire population (*i.e.* selected so that no two samples in the dataset are exactly genetically alike). Consequently,

samples rarely show complete genetic identity, even within a single window. Another, more flexible approach to uncovering the underlying trend of haplotypes was necessary in order to extract useful haplotype data from this dataset.

Initially, a script was developed to partition the 250K SNP data into small groups of aligned, contiguous SNPs, referred to from now as 'windows'. This was done in order to facilitate a sliding-window approach, in which sub-analyses of small sections of the data. For example, clustering analyses were combined to produce a more informative result than a single clustering analysis would provide on its own. Initially, the 216000 SNPs in the dataset were divided into 2167 windows of 100 SNPs; and later, the window size was decreased to 50 SNPs (4332 windows) to provide a better resolution of smaller haplotypes and greater accuracy in identification of larger haplotypes. SNPs in the 250K dataset showed an average separation of 1000 bp, though with a substantial range of variation from that figure. Given the average genomic separation required for breakdown of linkage disequilibrium noted by Kim *et al.* (Kim et al. 2007), a window size of 50 SNPs leaves little probability of a meaningful haplotype being falsely identified from the background linkage disequilibrium caused by simple linkage between nearby loci, while still giving a relatively fine resolution when identifying the bounds of genotypes that are conserved across multiple individuals. Since it is by definition impossible for a haplotype to extend across two chromosomes, care was taken to ensure that each window contained SNPs from only one chromosome.

Alignment of SNPs was not an issue in this analysis, due to the manner in which the 250K dataset was collected. The probes in the SNP microarray that were used to collect the 250K data contained SNP alleles at precisely known points along their lengths; this therefore marked the precise genomic locations of the SNP alleles, not by mathematical alignment tools, but by DNA hybridisation on the microarray. The upshot of this is that each SNP in the dataset is marked

with a specific locus number, (corresponding to its base-pair position along the length of the chromosome) and may safely be treated as aligned. While used in this project exclusively with a SNP dataset, this haplotype analysis was written to accommodate full genomic sequences too. Should this approach be applied to whole genome sequence data, however, it is recommended that the sequences are first transformed into a multiple alignment before splitting them into windows.

2.3.2 FIRST ATTEMPT

The first attempt at creating a tool to identify haplotypes from this windowed dataset involved two separate agglomerative clustering approaches. SNP sequences within each window were first clustered by forming groups based directly on the level of allelic similarity between samples. To accomplish this, clusters were assembled by carrying out pairwise comparisons between the short sequences of each window, each sequence compared SNP by SNP against every other, and a similarity score thus established by the simple expedient of counting the number of SNP alleles shared between them. Should that similarity score have crossed a preset threshold (*e.g.,* a similarity threshold of >=99%), the SNP sequence in question was added to an existing cluster if at least one of the sequences was found to belong to that cluster, or was formed into a new cluster if no such existing cluster was found. A given group of samples thus clustered within a single window was recorded, and henceforth referred to as a "hit". These single-window clusters were then extended by simply searching adjacent windows for clusters sharing more than a pre-set fraction of the accessions from the accessions shared between them. In this way, a set of 'hits' from adjacent windows could be strung together into a true haplotype, as a set of alleles spread across many adjacent loci on a chromosome and preserved across multiple individuals in the population. Since the accessions possessing this haplotype were also known, they could also thus be grouped together in a 'haplotype block'. The precise extent to which each accession in the block actually possessed the haplotype in question was also recorded, in the form of

'runs': a list of all the contiguous windows in which a given accession's SNP sequence clustered with the haplotype.

Although the exact details were different, some aspects of the method proposed by Zahiri *et al*. (2010) (see Chapter **2.3.1**) were already present. The similarity-based clustering of SNP sequences reflects the IsBlock function in its essential principle, in that it tested a region of the genome for its level of similarity, and returned a yes/no answer based on a clear and definable threshold. Also, the reconstruction of haplotypes through the extension of these local clusters had at least the same goal as the NeighbourBlock function, in that it aimed to extend a haplotype by incorporating data from adjacent loci. Crucially however, it did not once again call IsBlock on newly assembled haplotypes before accepting their extension as valid.

2.3.3 SECOND ATTEMPT

While demonstrating that a large number of potential haplotype blocks could be predicted, this approach had many flaws. When adding new SNP sequences into hit clusters, a requirement was set that the new sequence must have a greater-than-threshold level of similarity with at least 50% of the sequences already within a cluster. As a hit cluster grew larger, this simple acceptance condition could result in at least some of the SNP sequences within the cluster actually being much less similar to each other than the threshold level of similarity. For that reason, the method of joining single-window hit clusters into haplotypes spanning multiple windows sometimes continued to join adjacent hit clusters far beyond the point at which none of the accessions present at one end of the haplotype were found in the hit clusters constituting the other end. This was later somewhat mitigated by marking the centre-points of each accession's 'runs' of contiguous windows clustering with the haplotype, and applying a 1-dimensional K-means clustering. This sorted the runs themselves into groups; each of these groups was, from then on, treated as an independent haplotype block, thus resolving the problem of unfounded

extension. This could be regarded as correcting for false positives when constructing blocks.

Another flaw in this process was its speed in generating an output. Processing just the 136 accessions in the dataset that were gathered from the UK using this method took more than a week, despite several optimisation techniques being applied. Processing time for the whole range of accessions available in the 250K dataset was unacceptably large.

A second version of the haplotype-finder method was therefore created in an attempt to correct these flaws. A second implementation of IsBlock was written, which produced a distance matrix from the sets of aligned SNP sequences using the 'dnadist' clustering program (Felsenstein 2002), then resolved that matrix into a tree using UPGMA clustering (Sokal 1958). Clustered groups of accessions were derived from the tree by setting a threshold distance along the tree's branches, and taking branches falling below the threshold as cluster groups. Advantages of this method include: the use of dnadist provided much-improved measurements of the degree of similarity between SNP sequences within a window, producing a result substantially more quickly, and also allowing for subtle corrections such as adjusting similarity rates to account for differential rates of transition vs. transversion point mutations.

A second implementation of the NeighbourBlock function also extended 'hit' clusters into longer haplotypes by combining hit clusters from adjacent windows. In order to avoid the potential bias arising from the simple greedy algorithm of the first implementation that led to the excessive and unfounded extension of haplotypes in the first implementation, it was replaced with the Kuhn-Munkres algorithm (Munkres 1957). NeighbourBlock was called for each pair of cluster groups – i.e., all hits from a pair of adjacent windows. This created an assignment problem (representable as a bipartite graph), which the Kuhn-Munkres algorithm provides a near-optimal solution to. In order to prevent the algorithm making spurious pairings of clusters, another threshold

was set: pairs would only be expanded if the 'hit' cluster being added shared above a certain fraction of the base cluster's accessions. Typically this threshold was set at 50%, though later this was lowered. The complete set of SNP sequences of the new haplotype-cluster (the SNP sequence of all samples along every window within the potential new haplotype) was then passed once again to IsBlock and, if passing the similarity threshold, was accepted as a valid extension of the haplotype.

In this implementation of the haplotype-finder method, the SNP sequence data within each window was first partitioned into 'hit' clusters, as before. These were then joined, if possible, to hit clusters in adjacent windows by the Munkres-based implementation of NeighbourBlock. NeighbourBlock was set, at this point, to make several passes, in both directions along the genome, and to stop only once no further extensions to haplotypes could be made. The K-means clustering approach utilised at this point in the former implementation was reckoned unnecessary given this level of verification, and was discarded from this implementation.

The first attempt had shown that identifying haplotype blocks from dense SNP data was feasible, though its lack of speed had hindered its ability to analyse the dataset in full. The second attempt was designed to examine the entire, global dataset, and to do it more rigorously. In this, it succeeded; however, its requirement of three weeks to investigate and validate every possible extension of every retrieved haplotype remained unacceptable. Nonetheless, almost all of the elements required for a rapid and successful implementation were now in place.

2.3.4 THIRD ATTEMPT

The final, and most successful, implementation of the haplotype-finder method built mostly upon the second implementation, though re-incorporated some components developed in the first, and added further technical and scientific refinements.

The key realisation was that the ultimate purpose of the NeighbourBlock function might be approached from a different perspective. Instead of validating every haplotype extension before accepting it, a putative haplotype may be extended as far as possible first, and the validity of the extension checked (and amended, if necessary) later. As in the previous implementation, the first part of the analysis consisted of measuring the degree of similarity between the SNP sequences using dnadist, applying a hierarchical clustering to those similarity measurements, and creating groups based on a threshold of similarity between cluster branches. Likewise, the implementation of NeighbourBlock continued to use the Kuhn-Munkres algorithm to identify pairs of 'hit' clusters identified within each window, and thus to extend putative haplotypes. Unlike the second implementation, however, there was no minimum requirement of samples to be shared between adjacent 'hit' clusters for them to be provisionally accepted at this stage, besides, of course, that 'hits' in adjacent windows must share at least one sample. Additionally, the idea of making multiple passes over the genome was abandoned, and the original plan of making only a single pass was restored.

Finally, in order to counteract spurious block extensions, the K-means clustering functionality based on 'runs' of single samples' possession of a haplotype across a set of contiguous windows was restored, with one addition to prevent the opposite problem of spurious division of blocks. Prior to the application of the K-means algorithm, the distribution of run-centres along the overall length of the haplotype was subjected to a Shapiro-Wilk normality test (SHAPIRO & WILK 1965). If this returned a p-value indicating that the run-centres were normally distributed, the run was simply accepted as-is; should run-centres be normally distributed it is a plausible conclusion that the block is, as expected for reasons discussed more fully in Chapter **4.1.2**, centred around a single locus, perhaps one being driven towards fixation. However, if the normality test reported a likely absence of a normal distribution, the K-means clustering was then performed; a non-normal distribution suggests that at least

two loci within the set may be being drawn towards fixation. This solution aimed to avoid spurious division of putative haplotypes caused by the limitations of the K-means clustering algorithm in recognising datasets that are best represented by only a single cluster.

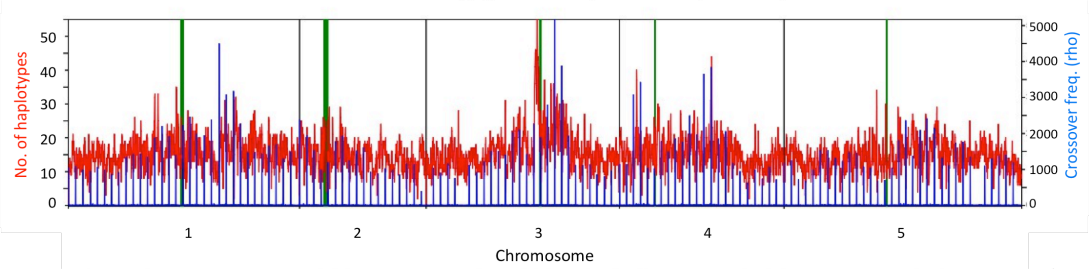


Figure 8 Haplotypes present within genomes of *Arabidopsis thaliana* sampled from UK populations. In order to carry out the analysis which determines the existence and extent of haplotypes within the sampled population (see Chapter **2.3.4**), high-density genotype data is divided into 'windows' of arbitrary length. In this project, a SNP dataset was divided into windows of 50 SNPs at adjacent loci. Advantages of haplotype discovery using this window-based approach are discussed in Chapter [**X.X**]. The number of haplotypes found in each window (**RED** series) across the genome is shown plotted against the meiotic crossover rate (**BLUE** series). (Data provided by Matthew Horton). Chromosome centromeric regions are marked in **GREEN**. While the crossover rate shows that recombination is strongly favoured at or near certain loci, the number of haplotypes remains relatively constant across the genome. ANCOVA analysis confirms no significant relationship between crossover rate and number of haplotypes; however, this is not surprising, as many of the genotypic clusters returned by the clustering approach can be attributed to population structure or chance similarity. Later analyses (i.e., SelectionFinder – see Chapter **4**) were developed to attempt to distinguish haplotypes subject to selective pressures from this background of haplotypes whose existence may be better explained by chance or other factors.

This final method combined the best aspects of the previous two implementations. Its run-time is short enough to analyse the entire 250K dataset in no more than a few hours; moreover, due to the addition of the normality test and K-means clustering to NeighbourBlock, the retrieved haplotype data generally matches the expected pattern of haplotype distributions within a population laid out in the introduction to this chapter. It should be noted that the use of dnadist as the primary means of quantifying the degree of similarity between genotypes means that this method could be applied to any large dataset of genomic variation, including complete genome sequence.

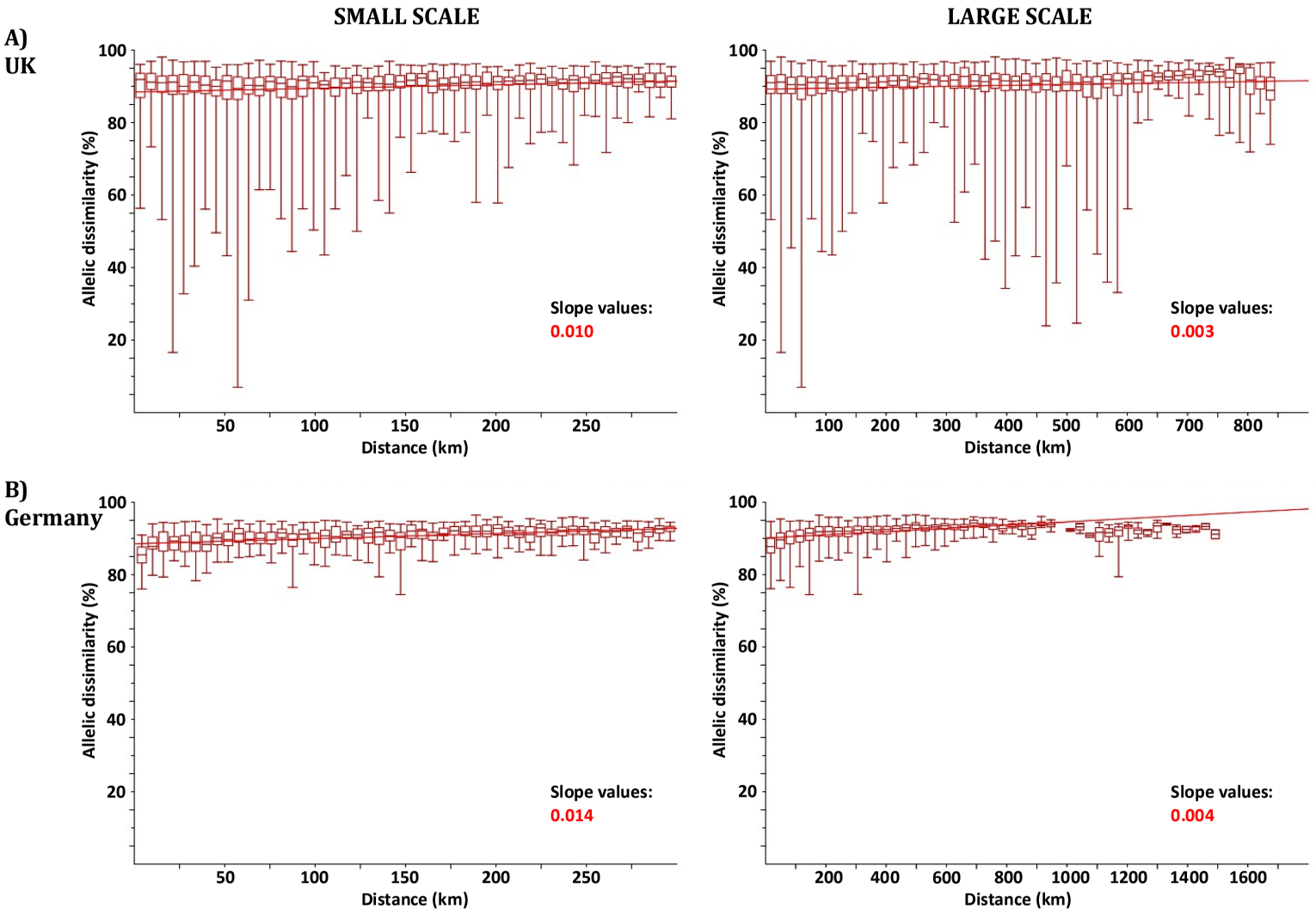
When this haplotype discovery method is applied to the 250K dataset, a large number of haplotypes (on the order of 110 000) are discovered. The number of haplotypes discovered at each window is plotted against the rate of meiotic crossovers at corresponding loci in Figure **8** (Data courtesy of Matthew

Horton). Despite the large amount of variation in crossover rates at different parts of the genome, the number of haplotypes found at any given locus remains relatively stable across almost most points in the genome. There appears to be no clear correlation between crossover rates and number of haplotypes.

2.3.5 SIMPLE POPULATION STRUCTURE ANALYSIS

Chapter **1.3.3** describes how work by Platt *et al.* (Platt et al. 2010) and Sharbel *et al.* (Sharbel et al. 2000) found that the *A. thaliana* population followed an 'isolation by distance' model, under which the genotypic similarity of individuals follows a trend of decreasing as geographic distance separating individuals increases. A simple analysis was carried out to determine whether the same observation was borne out by the 250K dataset – a dataset with a much greater density of genotypic observations than that used by Platt *et al.* (2010) and Sharbel *et al.* (2000)

This was sought in both the 'hit' data and the haplotype data generated by the haplotype finder, in a similar manner to that used by Platt *et al.* (2010): by simply plotting pairwise similarity of samples against the geographic distance separating their collection sites and applying a linear regression. Plotting a linear regression from this data shows that there is indeed a trend of decreasing genetic similarity over distance, as would be expected from a population following the isolation by distance model, across the UK and across the mainland populations in France and Germany (see Figure **9**). Plots in Figure **9** were derived from 'hit' data, but the same analysis applied to haplotype data yields essentially identical results, which are therefore not shown. 'Hit' data were chosen to represent this result since the cluster results from individual windows were a more similar data type to the set of individual SNP loci used by Platt *et al*. (2010), resulting in a more fair comparison between the two sets of observations.


SMALL SCALE
LARGE SCALE
A)
UK
B)
Germany
Allelic dissimilarity (%)
Distance (km)
Slope values:
0.010
0.003
0.014
0.004

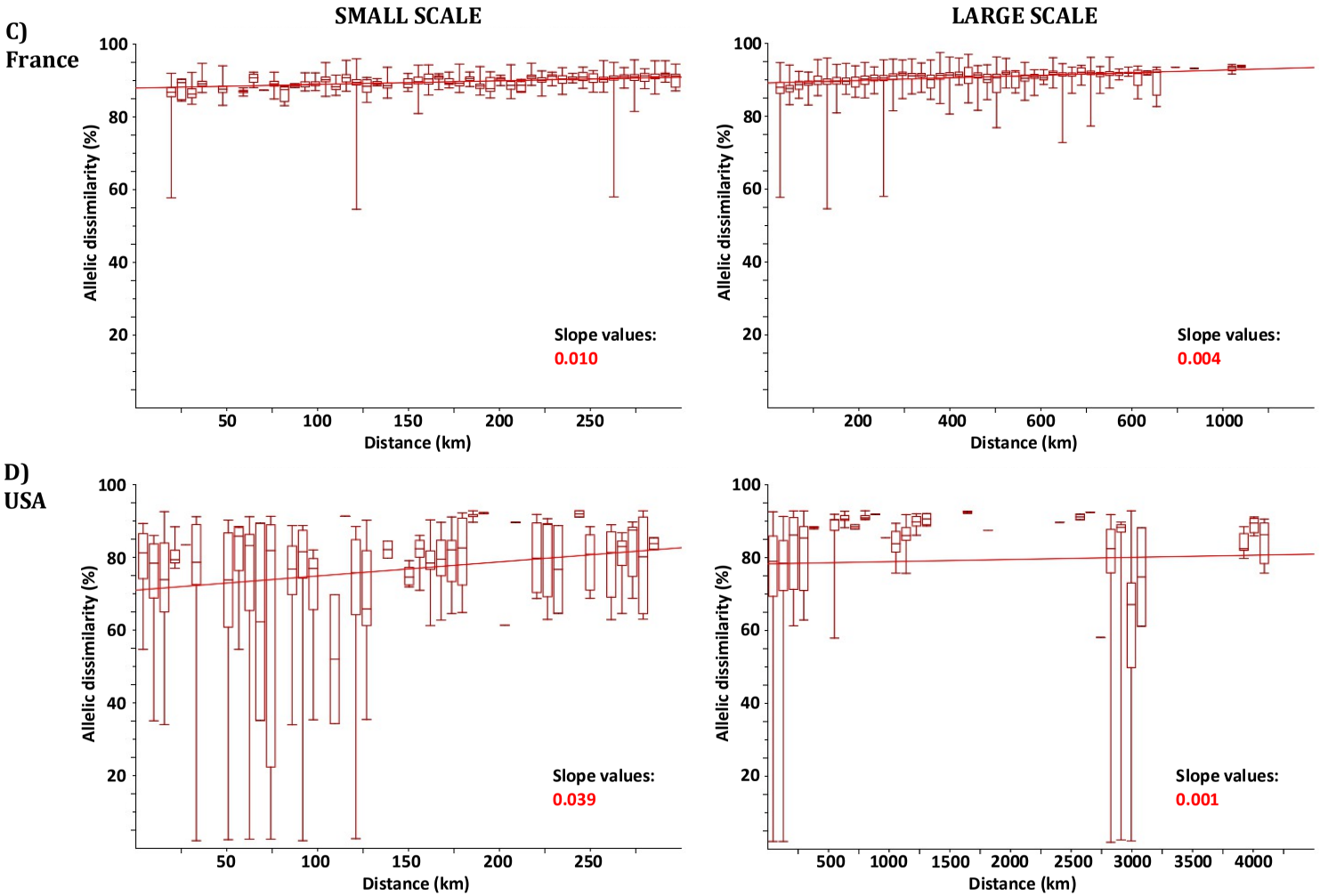

C)
France
SMALL SCALE
LARGE SCALE
Allelic dissimilarity (%)
Distance (km)
Slope values:
0.010
Slope values:
0.004
D)
USA
Slope values:
0.039
Slope values:
0.001

Figure 9 (Previous pages) "Isolation by Distance" analysis in different geographic samples of *Arabidopsis thaliana*. When the degree of genotypic similarity (or dissimilarity) between individual samples is quantified and plotted against the distance between the collection sites of those samples, a weak but consistent trend emerges: samples collected at more distant sites are more genetically different from each other. Plots show how the degree of allelic dissimilarity between samples at smaller (left) and larger (right) scales in **A**) the UK, **B**) Germany, **C**) France and **D**) North America. RED slope values indicate 'm' in standard linear regression equation y = mx + c (a linear regression is plotted for each data series). Genotypic dissimilarity between two samples was, in this project, quantified as the proportion of non-shared haplotypes divided by the total number of haplotypes across both samples (see Chapter **2.3.5** for details). This analysis replicates work by (Platt et al. 2010), and demonstrates that similar conclusions may be drawn from more dense data. Additionally, the degree that the data conforms to the plotted trend is informative: a population which closely follows the trend of isolation by distance was proposed by Platt *et al.* to emerge only in populations which have been established for a long duration. See Chapter **3** for full discussion of population structure and population establishment.

Within the mainland European populations, the trend of isolation by distance is closely followed – i.e., the majority of data points show a similar degree of genotypic similarity to that predicted by the regression. The isolation by distance trend within the UK, however, is a relatively loose one; there is a substantial range of deviation from the regression within some geographic distance intervals.

Platt *et al.* (2010) noted that the trend of isolation by distance appeared to be followed more closely, and over greater distances, in populations that have been established for a greater length of time (and vice versa). Specifically, Platt *et al.* (2010) noted that their observations of the population in the USA, which is known to have become established more recently than the Eurasian populations (Al-Shehbaz et al. 2006), were consistent with the expectations of the emergence of structure over time. While Platt *et al.* (2010) described this trend, no attempt was made to quantify the extent to which populations in given regions followed it, since Platt *et al.*'s (2010) analysis in this respect was focused on measuring the degree of intercontinental admixture rather than admixture within those populations.

An attempt to repeat and, additionally, to quantify this observation was made with the high-density data available to this project. Therefore, when regressions were plotted for genetic similarity vs. distance in sub-groups within the 250K data, the coefficient of determination (also known as R^2) of each

regression was also measured. R^2 quantifies the goodness of fit of the regression to the data; figures for the regression performed on each sample subset are shown in Table **2**. These figures show that Eurasian populations follow the trend of isolation by distance structure more closely in Eurasian populations than American populations, supporting Platt *et al.*'s (2010) conclusions.

Table 2: Goodness-of-fit of populations to Isolation By Distance trend

Subset of samples	R^2
UK	0.0077
France	0.1241
Germany	0.2184
USA	0.0017

2.4 DISCUSSION

2.4.1 HAPLOTYPE DISCOVERY METHOD

The method for haplotype discovery put forth in this chapter follows a sound method outlined by (Zahiri et al. 2010), and due to its reliance on long-established bioinformatic tools, is both fast and robust against a wide range of possibilities.

Its strengths lie in its speed and versatility. The method is capable of analysing a variety of data, including both SNP and sequence datasets; all that is required is that the data be partitioned into short segments along the length of the genome. Currently, this tool accepts input only in the specific format in which the 250K dataset is supplied, but since data is then supplied to the ubiquitous sequence analysis tool dnadist, modification to accept other common polymorphism data formats is trivial.

The arbitrary nature of the SNPs selected for each sliding window fails to take into account any variation in genetic distance between their loci or in the rate of meiotic crossovers occurring between them. Future work may refine this method by dividing the dataset into windows based upon map distances along chromosomes or recombination frequencies, as appropriate to the circumstances of the analysis.

The selection of SNPs for an analysis may be subjected to bias arising from several distinct sources: SNP discovery protocols may be biased towards the discovery of more common variants due to a small sample size; or towards finding variants located in parts of the genome that, due to biases in the sequencing process, are more deeply sequenced than other loci. The biggest advantage of the haplotyping method developed in this chapter is that, by defining clusters based on the consensus of many SNPs, ascertainment biases arising from unequal sequencing or other biases involved in the choice of SNPs comprising the dataset are not likely to result in significant alteration to the

overall trend revealed by clustering (see Chapter **2.3.1** and Figure **7**). Biases involving the disproportionate selection of samples possessing certain genotypes are accounted for simply by assigning those genotypes to appropriate haplotype blocks based upon the level of similarity of those genotypes, but this does not correct for any unequal selection of samples. This type of bias becomes a more significant potential concern in later chapters (see Chapter **4**), but ultimately is beyond the control of this project in any case.

A limitation of the method is that the output analysis is relatively basic compared to other analytical tools developed for similar purposes, such as ChromoPainter (Lawson et al. 2012). For example, linkage disequilibrium is not estimated directly; instead, its effects simply emerge through the process. Consequently, the analysis fails to detect haplotypes smaller than the set window size entirely, and is likely to struggle with those less than two windows long. This limits the ability of the analysis to examine old haplotypes, substantially degraded by extensive historical recombination. However, in the context of this project, direct analysis of haplotypes (rather than overall genetic similarity of samples) is carried out in order to examine very recent selection and migration. Analysis of historical migration and structure is therefore best carried out using the defined clusters within each subdivision of the genome – the 'hit' datasets.

This method was designed primarily to facilitate analyses in subsequent chapters. The haplotype and single-window cluster datasets created for this chapter enable extensive further analysis of the genetics and ecology of a wild population, and the code through which the method was implemented was also suitable for re-application in later work. For example, with the isolation by distance model acting as a baseline of expectations, it was possible to utilise the datasets produced with this method to carry out more advanced demographic analyses in Chapter 3 and to seek evidence of natural selection acting upon a wild population through selective sweeps in Chapter **4**.

2.4.2 POPULATION STRUCTURE ANALYSIS

Reports of the presence of a generalised population structure following a trend of decreased genotypic similarity of individuals with increased distance between sites of residence (isolation by distance) were re-examined with this high-density polymorphism dataset. Evidence from the regression analysis shows that the trend of isolation by distance does indeed consistently hold true across the European continent (as represented by the French and German populations), though not across the USA. Across Europe, linear regression shows that individuals at a great distance are, generally, only a few percent less genotypically similar to each other than individuals separated by smaller distances; however, this was also true of the samples analysed by Platt *et al*. (2010) The American population appears to show no consistent isolation by distance trend over large distances, but the US samples available in this dataset are relatively small in number for a large geographic area, and a larger and more consistent sampling effort may yet reveal a trend.

Platt *et al.* (2010) observed a less organised degree of isolation by distance in more recently emerged populations of *A. thaliana*, and from this proposed that isolation by distance emerges progressively across larger distances over time, and conforms more closely to the trend, as populations become more diverse and established. R^2 values show that the distance/similarity data for the UK population does not fit as well to its regression as the European populations, but fits better than the American population. .

The observed degree of isolation by distance appears to follow the predicted pattern of emergence over time in established populations. Given its prevalence in the European and French populations and absence in the American population, and that the Eurasian population is known to be ancient (Sharbel et al. 2000; Beck et al. 2008) and that the American population was established more recently (Al-Shehbaz et al. 2006), and Platt *et al*.'s (2010) conclusion that the isolation by distance structure becomes progressively established over

greater geographic ranges over time, it is reasonable to conclude from the degree of establishment of the phenomenon in the UK population that *A. thaliana* is a relatively recent arrival to the UK, in comparison to the duration of establishment of the European mainland population, but became established substantially prior to the founding of the American population. The high degree of differentiation observed at all scales across the UK, however, also shows that there has been significant gene flow between the UK and other populations, resulting in the introduction of a wide variety of genotypes. Structure analysis indicates that there has been relatively little admixture of genotypes since their respective migrations, however.

This is entirely consistent with the established model of Pleistocene population dynamics, in which populations rapidly moved to fill new habitats (including the British Isles) opened up by the retreat of the ice sheet. However, the isolation by distance model by itself offers little further detail on historical events.

This work sets a precedent allowing the use of degree of emergence of this type of population structure and differentiation of the population to estimate the time since founding of a sub-population, to create a baseline demographic model against which specific instances of gene flow may be compared, and to generate a set of expectations for the behaviour of haplotypes under selective neutrality. Given the simplicity and potential ubiquity of this haplotype analysis, it is likely that the tool developed in this chapter would prove as useful for investigating these matters in other species as it has here in *A. thaliana*. The specific analyses of isolation by distance population structure carried out here, however, are likely to be restricted in their usefulness to species sharing *A. thaliana*'s characteristics of an extensive geographic range and small individual dispersal distances.

CHAPTER 3: POPULATION HISTORY INFERENCE FROM GENOMIC DATA

3.1 INTRODUCTION

3.1.1 PLAN OF ATTACK

The first aim of this chapter was to test hypothesis **H2** (The UK population arose from a single source of founders) using more advanced ecological interpretation of the haplotype data acquired in the previous chapter. This involved an analysis identifying likely sources of the genetic diversity in the *A. thaliana* population in the UK, and further spatial analysis of both the UK-wide and Europe-wide population structure of the species. The results add directly to knowledge of the species and enhance subsequent analyses described in the next chapters.

The second aim was to test hypothesis **H3** (The UK population has been established for approximately 1000 years) by developing a means of estimating the number of generations since establishment of the *A. thaliana* population within a given area (in this case, the UK). This relied on the population genetic model of 'isolation by distance' that was described in the previous chapter. Time since founding was estimated by simulating the invasion of *A. thaliana* into the UK and its subsequent development of consistent patterns of isolation by distance.

3.1.2 POPULATION GENETICS & DEMOGRAPHIC CONCEPTS

The field of population genetics has grown over the many decades of its existence to incorporate numerous complex and powerful analyses, with great power to increase our knowledge and guide meaningful action in a variety of contexts. The basic concepts of the field (population structure, genetic drift, Hardy-Weinberg equilibrium and gene flow) have found numerous applications in studies of ecology and evolution. This introductory section will provide an

overview of the population genetic methods and concepts of relevance to this chapter.

Evolutionary research has historically utilised populations or species isolated on islands as 'natural laboratories', in which populations are free to adapt to their immediate local conditions without the counterbalancing effect of homogenising gene flow. In such a situation, local adaptation is likely to progress to the point of full speciation (Discussed in theory in (Kawecki & Ebert 2004), reported in wild plants in (Hall & Willis 2006). Given greater gene flow, however, the island population instead comes to represent an example of the invasion of the species into a new habitat range.

Typical population genetic analyses to resolve such situations involve a quantification of the level of differentiation (or, conversely, of gene flow) between isolated populations. The first aim of this chapter was, then, to measure the degree to which the UK *A. thaliana* population is differentiated from the populations on mainland Europe, and in doing so, to identify the sources of genetic variation present within the population. However, this is complicated somewhat by the nature of the general population structure, as shall be explained later (see Chapter **3.1.3**). Instead, background necessary for a comparison between populations resident within different habitat types across the UK will be addressed first.

When a population may be divided into several groups based upon geographical separation, population genetics offers many tools for quantifying the degree of differentiation between them. The most commonly used of these is F_{ST}, the fixation index, which is estimated by comparison of the degree of genetic variation within sub-populations to the degree of variation between sub-populations (Wright 1950). It ranges from 0 (no hindrance whatsoever to gene flow between sub-populations, as described by Hardy-Weinberg equilibrium) to 1 (complete isolation of populations). If the simplifying assumption of sub-populations restricted to defined areas is made, then F_{ST} can

be used to estimate the number of migrants between areas per generation (Crow 2010).

Tajima's D is another informative statistic (Tajima 1989), itself derived from a comparison of two other statistics: the number of segregating sites (i.e., those that are polymorphic) in the genomes which constitute a species' gene pool, and the mean number of pairwise genetic differences between individuals. When applied to a population possessing no structure and under no pressures of selection, this statistic will not significantly differ from zero. The statistic is used to infer the type of selection acting on a given segment of the genome; negative values indicate the presence of directional selection via selective sweeps, while positive values indicate the presence of balancing selection. As with other means of detecting selection, however, it is also susceptible to influence from population structure. Tajima's D has been shown to be generally negative across the Arabidopsis genome, most likely due to recent population growth (Nordborg et al. 2005).

Prior analyses of the wild *A. thaliana* population have, as shown in Chapter 1, frequently demonstrated that the population structure follows the isolation by distance model. While this observation is undoubtedly useful, it does little on its own to elaborate on specific details of gene flow between areas or habitats, or to indicate the historical demographics of the species including its ultimate origin. In particular, prior work has not looked in detail at the sources of genetic variation within the *A. thaliana* population inhabiting the UK, nor specifically set out to measure the extent to which this island population is differentiated from populations in mainland Europe.

Moreover, this model raises difficulties in analysis using the most common means of estimating population genetic statics such as F_{ST}. Within the UK, data regarding the distinct habitat types from which samples were collected is available to this project, defining several groups suitable for such a comparison. Outside of the UK, however, the natural distribution and population structure of

A. thaliana generally does not form a set of groups that would make for an obvious application of such techniques, and further data on the habitat types at which samples were collected is unavailable to this project.

Other analyses must be developed to advance our knowledge of historical demographics of *A. thaliana*. For example, a more advanced avenue of population genetic analysis uses coalescent theory. Typical coalescent analysis aims to identify the most likely sequence of historical changes in alleles leading back to their most recent common ancestor. Knowledge of the average rate of mutation then enables an estimate of the amount of time that has passed since the divergence of the alleles. However, confidence in findings using this type of analysis still ultimately requires a good understanding of the demographics of the population, in order that a sound prediction of historical changes is made. On a national and continental scale, this knowledge is currently unavailable, and due to extensive admixture (see next sub-chapter), is likely to remain elusive for the foreseeable future.

3.1.3 DEMOGRAPHIC HISTORY OF A. THALIANA

Over the past 15 years, a large number of studies have built up a comprehensive model of general demographic history of *A. thaliana*. Sharbel *et al.* (Sharbel et al. 2000) initially demonstrated that *A. thaliana* populations do not display sufficient reproductive isolation to form a 'phylogeny of ecotypes', but instead show a large degree of historical admixture. Based upon measurements of genetic similarity across AFLP loci, they concluded that genotypes present in Western Europe today primarily arose from a population inhabiting the southern Iberian Peninsula while the Pleistocene ice sheet covered much of northern Europe. As the planet warmed and the ice began to retreat, populations from refugia in the Iberian Peninsula and Caucasus regions expanded to colonise newly available habitats opened up by the melting of the ice. Eventually, the ranges of these populations expanded to the point of reconnection, leading to distinct patterns of admixture in eastern

Europe/Eurasia known as 'suture zones'. Since similar demographic patterns have been observed across a large number of species, this model has come to be known as the 'Pleistocene paradigm' (Hewitt 1999).

Beck *et al.* (Beck et al. 2008) later refined this model with a prediction that *A. thaliana* most likely first arose in the Caucasus and initially spread westwards across Europe prior to the encroachment of the ice sheet, before returning westward to meet populations from Asian refugia as the ice retreated. François *et al.* (François et al. 2008), however, proposed a very different and conflicting interpretation – that instead of advancing eastward from Western refugia as the ice retreated, *A. thaliana* populations from Eurasia rapidly spread westward, assimilating and replacing western populations into their wave of invasive advance. François *et al.* (2008) state that this relentless advance may have been a natural event, but speculate that *A. thaliana* was inadvertently spread through the growth of Neolithic farming practices across Europe.

Throughout all of these analyses, the UK population has remained relatively under-studied, in that no analysis has set out specifically to investigate its demographics. This is unfortunate, as the UK population presents intriguing scientific questions. As demonstrated in the previous chapter, there is reason to believe that the UK population was established substantially more recently than populations at comparable latitudes. It is not immediately obvious why this is so; in the late Pleistocene, the British Isles were connected to mainland Europe via a land bridge, and many other species are known to have colonised the UK. Moreover, it is not known exactly how long ago the UK population was founded, where the founding individuals were sourced from, or why the founding occurred at that time rather in concert with the expansion across the rest of the mainland.

This chapter aimed to gather data to help in resolving this question; first through a Europe-wide analysis of genotypic similarity to identify the source (or sources) of variation within the UK population, and secondly through the

development of a tool to infer the age of such an island population from its degree of emergence of population structure.

3.1.4 DATA REQUIREMENTS: GENE FLOW, STRUCTURE, DISPERSAL AND RECOMBINATION

In order to carry out the analyses described in this chapter, several key pieces of data are required. Gene flow within the species' range must be known; this was accounted for with a simple implementation designed to reflect the isolation by distance population structure already extensively discussed. The relative importance of physical mechanisms of gene flow must also be accounted for in order to produce an accurate model. Bakker *et al.* (Bakker et al. 2006) proposed that despite the low outcrossing rate of the species, pollen dispersal is the major mechanism of gene flow across the *A. thaliana* habitable range. To reflect this, pollen dispersal was simulated over much greater distances than seed dispersal. Knowledge of meiotic crossover frequency per chromosome is also required; Giraut *et al.* (Giraut et al. 2011) reported a crossover rate of 1-3 chiasma per bivalent.

3.2 MATERIALS AND METHODS

3.2.1 PRINCIPAL COORDINATE ANALYSIS, STRUCTURE AND CLUSTERING

A principal coordinate analysis (PCA; also known as multidimensional scaling – MDS) was carried out on the set of genotypes identified in the previous chapter. A pairwise distance matrix was prepared, in which the degree of genetic difference between samples was quantified by counting either the number of 'windows' in which the two samples clustered together and dividing by the total number of windows; or, similarly, by counting the number of haplotypes possessed by the two samples and dividing by the total number of haplotypes present in both.

A script to prepare this matrix was written in Perl, utilising the modules List::Compare (for ease of tallying shared/non-shared points of comparison) and Math::NumberCruncher (for common mathematical functions). The principal coordinate analysis was then carried out in R. The two major principal components were exported from R and plotted with a second Perl script, utilising the modules Image::Magick (for construction of graphics) and Chart::Math::Axis (for axis scaling). Clusters within the UK population were identified from this data by marking groupings in the two major principal components.

The habitat types in which the samples belonging to each of these clusters were found were compared against a null hypothesis ("Samples belonging to the cluster will be distributed across habitat types at the same ratios as the entire UK population") using Chi-square tests. Geographic ranges of each of these genotypic clusters was shown by finding the convex hull of the sample collection sites attributable to a genotype cluster, and by plotting that hull over a map with an R script utilising the 'maps' package.

Clustering by PCA was also supported by a second clustering analysis, using the population genetics analysis tool Structure (Pritchard et al. 2000). Due to the

high rate of self-fertilisation in *A. thaliana*, any dataset incorporating diploid genotypes risked violating the assumption of independence of loci inherent to the analysis (Falush et al. 2003; Beck et al. 2008). These previous analyses have avoided this issue by simply selecting one allele from a heterozygous pair, in order to produce an effectively haploid genotype; and since the 250K dataset used in this project was already supplied in a haploid state, it was necessary to employ this approach in any case.

Upon the completion of the PopAger simulation, genotype data from the simulated population was itself subjected to both PCA and Structure-based clustering analyses, in order to verify that the simulated population accurately represented the overall distribution of genotypes and relatedness observed in the wild population.

Finally, pairs of accessions possessing an unusual degree of genotypic similarity over distance were identified using a simple nonparametric likelihood plot of measurements of similarity between accessions separated by comparable distances. This was implemented using a Perl script.

3.2.2 POPAGER TOOL

The PopAger tool was created in Perl, and incorporates a number of modules already utilised in other areas of this project, accessible from the Perl repository CPAN (see Chapter **2.2.1**. Geographic distances between sample collection sites were calculated using the Haversine method (Robusto 1957) using module GIS::Distance. The alternate option of specifying distance between samples (and thus the degree of migratory and reproductive interchange) via output from Structure clustering (Pritchard et al. 2000) was implemented, but was not used in this project. Any situation requiring differences or similarities between two lists was handled by module List::Compare. Various common mathematical functions were carried out by module Math::NumberCruncher.

Number of crossovers in any given simulated outcrossing event was given a realistic variation in range (a normal distribution centred around the supplied mean value of 2 per chromosome (a figure derived from (Giraut et al. 2011)) using module Math::Random::OO:Normal. The descendent genotypes were assembled from the chosen pattern of crossover points using module IntervalTree. Weighted random choices, including the sites to disperse newly-created seeds to, and the loci at which crossovers were to occur in an outcrossing event, were carried out using module List::Util::WeightedChoice.

Linear regression was applied to datasets generated from both wild and simulated populations using the module Statistics::LineFit. Comparison of regression trends from these two datasets was achieved through an analysis of covariance (ANCOVA), implemented through the module Statistics::Distributions::Ancova.

Data detailing the variation in crossover rates across the genome was kindly given by M. Horton.

3.3 RESULTS

3.3.1 PRINCIPAL COORDINATE ANALYSIS, STRUCTURE ANALYSIS AND CLUSTERING

An attempt was made to investigate the migratory history of the UK population by searching the global set of haplotype data for genotypes similar to those encountered within the UK. Once clusters of genotypes were established within the UK, a subsequent analysis was performed to test whether some of those genotype clusters were disproportionately encountered in particular habitat types, a situation that could be suggestive of those genotypes being adapted to that habitat type.

In addition to the spatial data of each sample's collection site, some details of the immediate habitat were collected and included in the 250K dataset. Habitats were classified into four types: 'wall/rocky outcrop' sites, characterised by growth of plants in communities clinging to a hard substrate of either natural or human origin, and likely remaining undisturbed for long periods (>20 years); 'garden' sites, characterised by softer and richer substrates and a potentially higher incidence of human disturbance; 'railway' verge and ballast sites, characterised by a low incidence of human disturbance, but perhaps most greatly favouring long-distance dispersal; and 'other', characterised by a habitat that falls into none of the above categories (see Appendix **4** for table).

The null hypothesis for this analysis was that genotypes would be distributed proportionately across all habitat types. This hypothesis was tested using a set of chi-squared tests, with one test applied to the observed frequencies of samples from each habitat within each cluster of genotypes. Results from this analysis (Table **3**) show that the null hypothesis was rejected for the UK-Scandinavian cluster of genotypes, with the results showing that these genotypes were encountered almost exclusively in 'wall/rocky outcrop'

habitats. All other genotype clusters, however, show no firm evidence of favouring one habitat over any other.

Analysis of F_{ST} was carried out between the populations within each habitat type, in order to quantify the degree of differentiation between the populations in these habitats. 1000 loci were selected at random for the analysis. F_{ST} was estimated in the standard manner:

$$\mathbf{F_{ST} = (B - W) / B}$$

in which B is the average number of pairwise differences between populations, and W the same within populations. Estimates of F_{ST} between habitats within the UK are shown in Table **4**.

A similar analysis between the UK-wide and mainland populations was considered, but ultimately the option to carry it out was declined; while the UK samples form an obvious group for comparison, the mainland populations do not naturally fall into such easily distinguishable groups.

Table 3 Chi-square test results of clusters vs. habitat P-values indicate no significant deviations from chance expectations of distribution of samples across habitats except for the UK-Scandinavian cluster, which was found almost exclusively in wall/rocky outcrop habitats.

CLUSTER	UK-Scandinavian	UK-German	UK-Iberian-French	UK-French	UK-only
Chi-square	20.696	0.719	3.390	2.933	3.139
DOF	8	4	37	46	34
P-value	**<0.01**	**>0.25**	**>0.25**	**>0.25**	**>0.25**

The geographic ranges of each cluster of genotypes were then explored. The sampling locations of the accessions comprising the UK-Scandinavian, UK-German, UK-US-Iberian-French, UK-French and UK-only clusters derived from the

Table 4 Values for F_{ST} between habitats Low values of F_{ST} indicate little obstruction to gene flow between habitats.

Comparison	F_{ST}
Wall/Outcrop vs. Garden	0.023
Wall/Outcrop vs. Railway	0.019
Wall/Outcrop vs. Other	0.029
Garden vs. Railway	0.000
Garden vs. Other	0.002
Railway vs. Other	0.010

principal coordinate analysis (Figure **10**) were encircled on a map using convex hulls (Figure **11**). This shows the parts of the UK over which each set of genotypes is dispersed. The UK-Scandinavian cluster appears to be restricted to the northwest of the British Isles, while all other clusters are distributed across almost the whole landmass of the UK.

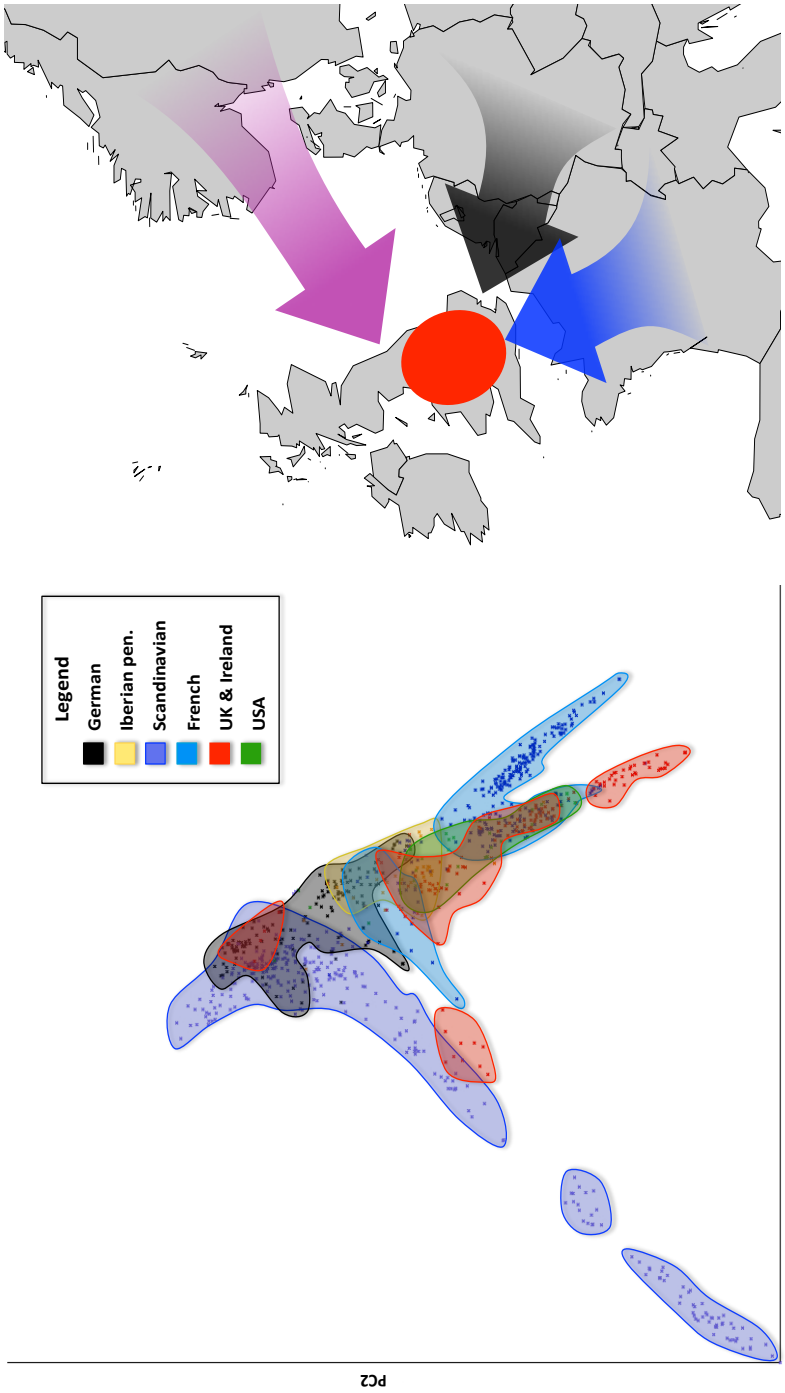
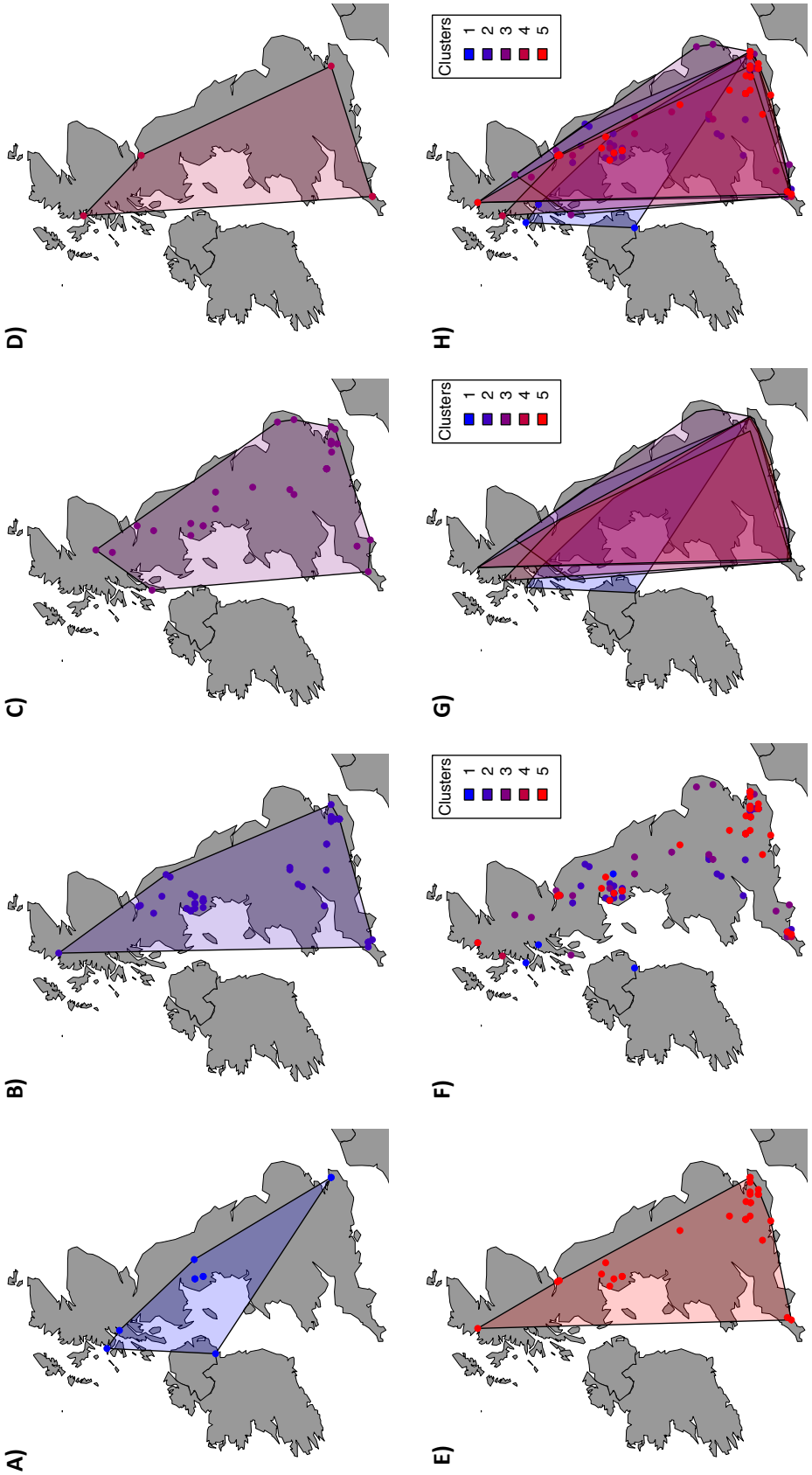


Figure 10 (Previous page) Principal Coordinate Analysis (PCA) of haplotypes in a global sample of *Arabidopsis thaliana* and inferred historical movements. Data comprising a pairwise quantification of genetic differentiation across all European samples represented in the 250K dataset was subjected to a principal coordinate analysis. Clusters are enclosed with colours indicating general collection zones. Principal components 1 and 2 explain 30% of observed variation. Accessions collected in the UK (**RED** crosses in the left-hand graph) separate into four main groups: three clusters share general genotypic similarity with samples from other regions across Europe, and one additional cluster appears to be unique to the UK. This genotypic similarity suggests that gene flow or migration has occurred between the UK and the areas from which the mainland samples were collected (represented in the right-hand figure). Since this project has found evidence that the UK population was likely established more recently than the mainland populations (Figure **9**), this means that these migration/ gene flow events are more likely to have been directed into the UK from mainland populations than vice versa. However, one other cluster of UK genotypes is relatively distinct from others represented in the dataset. This was initially believed to represent a population which became established and was able to become differentiated from the mainland populations due to a lack of gene flow, but further experimentation demonstrated that this cluster was capable of being generated without any provision for early founder events or isolation. For full discussion, see Chapter **3.4.1**.

Figure 11 (Next page) Geographic distributions of genotypic clusters across the UK It is possible that some of the genotypic clusters identified by principal component analysis (see Figure **10** for cluster plotting, and Appendix **5.1** for lists of samples comprising each cluster) are adapted to specific habitat conditions other than those represented by the set of 4 possibilities represented in the 250K dataset. For example, a reasonable hypothesis is that the cluster which appears to be derived from the Scandinavian population may be better adapted to general climatic conditions found at latitudes closer to those at which the source population occurs, and that samples possessing that genotype would therefore be found more commonly – or exclusively – towards the north. The collection points of the samples comprising each cluster were plotted on a map, and a convex hull was drawn to demark the area known to be inhabited by individuals bearing that class of genotype. Maps **A)** to **E)** show the collection sites of the UK-Scandinavian, UK-French, UK-Iberian/French, UK-German and UK-only genotype clusters, respectively. Maps **F)** and **G)** show the complete set of collection sites and convex hulls overlaid, and map **H)** shows all available data combined into a single plot. All genotypic clusters appear to be distributed widely across the UK, except for the UK-Scandinavian cluster, which is not represented in the south-west of the UK. Similarly to the conclusion drawn from the analysis described in Chapter **3.3.1** and Table **3**, which revealed that samples in the UK-Scandinavian genotype cluster are disproportionately found in wall/ rocky outcrop-type habitats, this suggests that those genotypes may also be better adapted to environmental conditions encountered in the north of the UK than in the south. However, as with the other analysis, this conclusion is based on a very small number of samples, and should be treated as a hypothesis to be tested by further experiment.


A)
B)
C)
D)
E)
F)
G)
H)
Clusters
1
2
3
4
5

Clustering was also carried out using the population genetics analysis tool Structure (Pritchard et al. 2000). The clusters produced by Structure analysis at k=5 were in generally good agreement with those produced by the principal coordinate analysis, with the majority of samples being placed into the same clusters by both approaches. Clusters produced by Structure are shown in Appendix **5.2**, and admixture between samples is shown in Figure **12**.

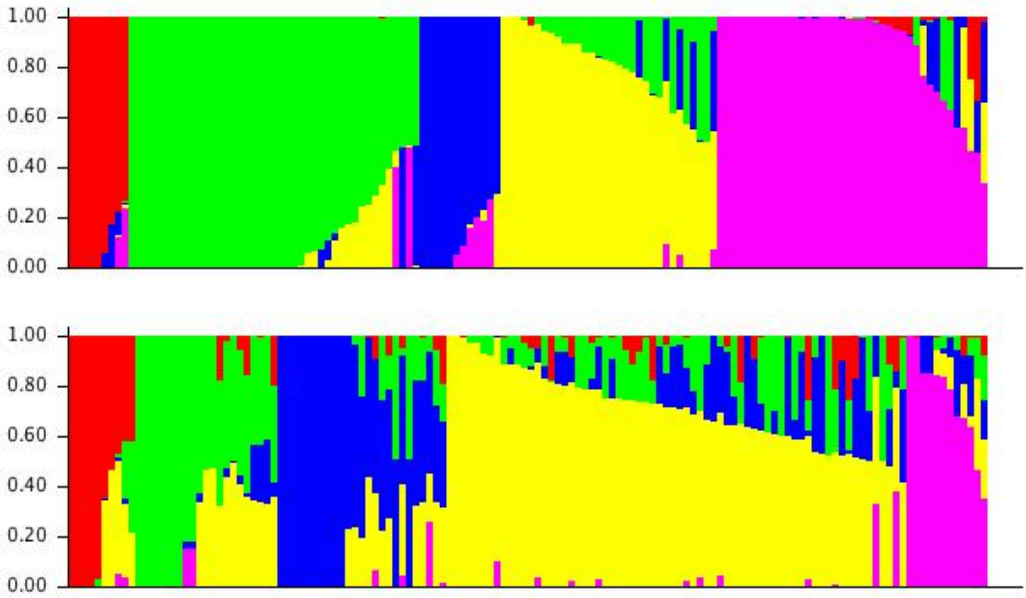


Figure 12 Structure analysis of *Arabidopsis thaliana* genotypes sampled from UK populations with sampling location set to reflect habitat type. K=5 clusters provided the best fit. The upper plot shows the frequencies of and degree of admixture between genotype clusters in the UK samples within the 250K dataset, while the lower plot shows the same for UK genotypes produced by the PopAger simulation (see Chapter **3.3.6**). This shows that although the PopAger simulation does not perfectly recreate the frequencies of genotype clusters across the UK, the simulated population possesses a degree of admixture comparable to that observed from the wild population.

Anastasio *et al.* (Anastasio et al. 2011) identified 286 accessions from a set of almost 6000 obtained from stock centre seed banks as potentially mislabelled or contaminated. In order to investigate the possibility that some of these accessions may be attributable to long-distance dispersal as opposed to seed stock contamination or mislabeling, pairs of accessions showing the greatest degree of genotypic similarity for their separation distance class were extracted. Two thresholds were set for extraction of pairs: p=0.05, and p=0.01.

Figure **13** is a summary plot showing the distribution of these extracted pairs. Pair data from both analyses is shown in Appendix **1**.

Figure 13 Paths of greatest dispersal likelihood across the UK based on samples showing the highest degree of genotypic similarity for their distance class. Lines mark pairs of accessions with a degree of similarity departing from the trend at p=0.05 (**LEFT**) and p=0.01 (**RIGHT**). **RED** points mark samples from the 250K dataset. This data may be employed in future experiments designed to assess the extent to which inadvertent human action plays a part in the dispersal of wild plant species.

Data containing a quantification of genotypic dissimilarity and distance between pairs of accessions (*i.e.,* the same as that used to investigate population structure in the previous chapter) was separated into classes based on distance, and the most genotypically similar accessions within each class were extracted. Accessions possessing either an unexpectedly high degree of genotypic similarity with other accessions far outside the normal geographic range of their haplogroup, or no means of corroborating the genotype with other regional samples, were placed on a 'red-list'. Data from this project (Appendix **1**) shows that some of these apparent mislabelling/contamination events may instead be attributable to gene flow caused by human action, and that this matter is worthy of further investigation.

3.3.2 EFFECTIVE POPULATION SIZE AND DISPERSAL PARAMETER SCALING

In the PopAger and SelectionFinder simulations (Chapters **3.3.4** and **4.3.2** respectively), the isolation by distance population structure is caused to emerge as a consequence of the dispersal of seeds and pollen from the habitation sites of their parents (see Chapter **2.4.2** and Figure **14**). The likelihood of dispersal from one site to another is calculated from an exponential decay function (i.e., dispersal likelihood is modeled on a Pareto distribution (Arnold 1983). In order for seeds to disperse an average of 1 metre, for example, the exponent of the decay function would be set in order that the probability of a seed dispersing to a given site 1 metre away is half that of the seed remaining at the same site of its parent, twice that of a site 2 metres away, and 4 times that of a site 4 metres away. See Figure **14** for a cartoon representation of seed dispersal likelihood.

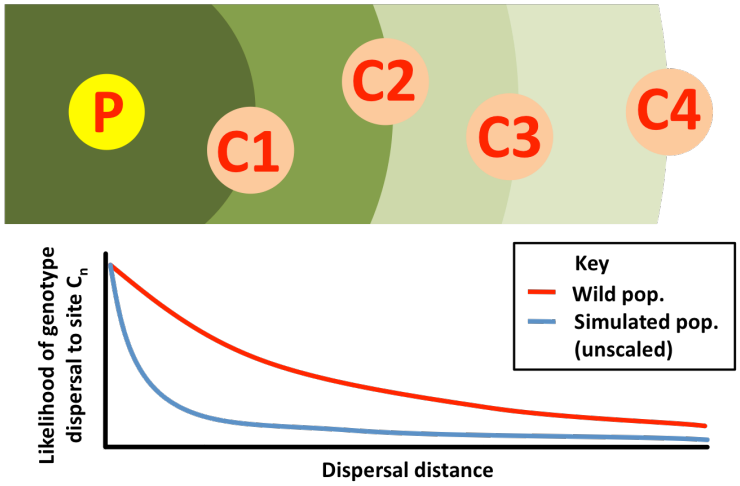


Figure 14 The requirement for seed/pollen dispersal scaling When the PopAger and SelectionFinder simulations are supplied with seed and pollen dispersal parameters matching the values for real plants, the volume of gene flow between sites represented in the simulation is small, due to the relatively large distances between collection site represented in the 250K dataset compared to the average dispersal distance of an individual seed and the small number of individuals able to be represented in the simulation. This leads to a simulated population of small, isolated groups that does not reflect the degree of genotypic interchange observed between sites in the wild population (see Chapter **2.4.2**). This discrepancy is likely attributable to the larger number of habitable sites available to the wild population, which would allow gene flow between distant sites to take place through successive short transfers between nearby sites. Since these 'stepping stone' sites are not represented in the simulation, the dispersal distances of simulated seeds and pollen must be increased in order to bring the volume of gene flow between the small number of sites represented in the 250K dataset to a level comparable with that observed in the wild. See Chapter **3.3.2** for calculations used to scale dispersal likelihood to better reflect gene flow in the wild population, and Chapter **3.4.2** for discussion of the consequences of the lack of 'stepping stone' habitation sites and dispersal scaling on conclusions drawn from PopAger analysis.

Before the PopAger tool can be utilised, we must first establish sensible dispersal parameters for simulated seeds and pollen, in order that the exponent values for seed and pollen dispersal can be set. In an ideal situation, the number of individuals represented in the PopAger simulation (and also in the SelectionFinder simulation – see Chapter **4**) would approximate the number of individuals in the wild population. Unfortunately, this is impossible within the scope of this project, due to the limitations of both the 250K dataset and of the amount of compute power available. In order to avoid causing issues involving the interactions of sub-populations within a metapopulation – i.e., excessive accumulation of genotypic diversity within habitable sites, and therefore a failure to accurately represent the highly homogeneous sub-populations (known as 'accessions') that *A. thaliana* typically exists in (see Chapter **1.3.3**) – the number of individuals simulated was restricted to the number of UK samples in the 250K dataset multiplied by a small number (typically 10) chosen to represent a typical small stand of *A. thaliana* (Holub 2012, *pers. comm.*.). This number was also near the upper limit of the number of individuals that could be simulated using the approaches discussed in Chapters 3 and 4, given the amount of compute resources available to this project.

When seed and pollen dispersal figures approximating those of plants in the wild (discussed the end of this sub-chapter) are supplied to PopAger (see Chapter **3.3.4**) and the simulation is allowed to run, essentially no gene flow occurs between plants represented at the sampling sites represented in the 250K dataset. This is not consistent with the F_{ST} values reported in Table **4** (representing F_{ST} values approaching 1, as opposed to the values approaching 0 observed in the wild population), and therefore a simulation supplied with those parameters is not an accurate representation of the real population and cannot be relied upon as a means of drawing accurate conclusions concerning the real population. The discrepancy in the volume of gene flow between the simulated and real population is likely caused by the significantly smaller number of individuals and habitable sites in the simulation than in the real

world; a larger population spread across a larger number of sites offers more opportunities for a genotype to be dispersed from one area to another.

Consequently, in order to represent the volume of gene flow occurring across the UK in the wild population, the simulated individuals were required to disperse a greater distance. By causing simulated seeds and pollen to disperse greater distances, a greater volume of gene flow between sites represented in the simulation was achieved. The equations in the following pages describe how the mean dispersal distances of seeds and pollen were adjusted in order to bring the simulated interchange of genotypes between sites up to a level comparable with that observed in the real population (see Figure **14**). While this adjustment proved effective in the case of this project (PopAger simulation produced a set of genotypes exhibiting a similar degree of isolation by distance population structure to that observed from the wild population; see Chapter **3.3.4**), PopAger and SelectionFinder analyses should not be used to draw conclusions in any situation in which *a posteriori* analysis of the structure of the simulated population does not return similar results to those of population structure analyses on the wild population.

Chapter **3.4.2** further discusses the possible effects of the scaling of this parameter on the reliability of conclusions drawn from PopAger analysis.

The first step in deriving these figures was to get an estimate of the effective population size (EPS) for the UK *A. thaliana* population. This was done using the standard equation (taken from (Fu 1994)):

$$\theta = 4N_e\mu$$

which rearranges to

$$N_e = \theta / (4\ \mu)$$

where θ is the population mutation rate, N_e is the EPS, and μ is the mutation rate per base pair per generation. The value for μ was taken from Lynch and Conery (Lynch & Conery 2003).

θ was calculated by taking all of the SNP allele differences between each pair of UK accessions and dividing by the number of comparisons, effectively giving the mean number of SNP allele differences between any two UK accessions. It is possible to refine this figure, however; while HapMap data fails to capture the full range of genetic variation between two given samples, that information is now available from the 1001 Genomes project. Two UK accessions were selected at random, and the difference between the number of SNP allele differences and genomic data differences used to calculate a 'correction factor', like so:

$$\mathbf{Cf = D_g / D_s}$$

where Cf is the correction factor, D_g is the number of allelic differences between the two accessions measured from genomic data, and D_s is the number of allelic differences measured from SNP data.

The correction factor may then be applied to the effective population size equation like so:

$$\mathbf{N_e = \theta / (4 \mu Cf)}$$

The addition of the correction factor reduced the estimate of the EPS by around 2 orders of magnitude, to a final value of 4,851,000. When applied to the PopAger simulation (and, later, the SelectionFinder simulation), this correction factor causes the small number of individual plants represented per generation at the small number of sampling sites available to the analysis to disperse to a much greater extent – recreating the level of dispersal expected between distant sites in the larger and more scattered extant population.

A simulation run without the introduction of this correction factor to dispersal likelihood calculations results in small populations with essentially no movement of genotypes between sites: a simulated F_{ST} of ~1, which – as shown by F_{ST} measurements of the wild population in Table **4** – bears no relation to reality. See Figure **14** for a graphical representation of the function of the dispersal correction factor in matching simulated gene flow to that observed from the wild population.

This value may now be used to begin to calculate dispersal parameters for simulated plants. In practical terms, the simulation of 4.8 million plants remains impractical given current computational resources. However, a smaller simulated population may be taken as representative of a larger one if dispersal and outcrossing parameters within the simulated population are scaled up appropriately in order to match with the rates of gene flow and migration observed in the larger real-world population. An appropriate set of such parameters enables the simulation of large populations over many thousands of generations in a reasonable timeframe using only modest computational resources. Given a number of plants to be simulated (dictated by the number of accessions collected and by the number of each the experimenter wishes to simulate), the real and simulated populations balance like so:

$$\mathbf{P_r D_r = P_s D_s}$$

where P_r is the number of real-world plants (i.e., the effective population size), D_r is the median dispersal distance for real seeds (which may be imagined as a 'seed dispersal distance half-life'), P_s is the number of simulated plants, and D_s is the median dispersal distance for simulated seeds.

Since we want to extract the value D_s, this rearranges to

$$\mathbf{D_s = (P_r D_r) / P_s}$$

Additionally, the fact that not all seeds disperse to a suitable growing site (or otherwise fail to grow to maturity and produce seeds of their own) must be accounted for here, using factors termed 'seed survival values'. When running the simulation under a neutral model, the size of the population will not significantly change between generations, since the number of simulated plants is hard-capped, and each generation's plants invariably produce more than one seed each. However, the situation in the real world is more complex, in that the chance of a seed landing at a suitable site and reaching maturity is lower than that of a simulated plant. These adjustments allow that difference to be accounted for, like so:

$$\mathbf{D_s = (P_r D_r S_r) / (P_s S_s)}$$

where S_r is the seed survival value for the real population, and S_s is the same for the simulated population. When using this tool under neutral conditions to estimate population age, S_s was set to $1/P_s$ in order to reflect the inherent stability of the simulated population.

(Note: should this tool later be used to model selective effects acting upon a population, the seed survival values will in all likelihood still need to be used, since the simulated population is still unlikely to show comparable rates of seed survival to maturity without this correction being applied).

While this establishes the principle, it must be remembered that *A. thaliana* generally does not occur in isolation, but in small, genetically homogeneous or near homogeneous stands in single habitable sites (described as 'accessions'). This is reflected in the simulation tool, which gives the option of allowing multiple plants to colonise any given sampling site using the 'number of plants per site' setting. Since the simulation has been set up to use the likelihood of dispersal between sampling sites, the values P_r and P_s must be altered to account for this by altering them to reflect the number of potential sites rather than the number of plants:

$$D_s = ((P_r / N_r) D_r S_r) / ((P_s / N_s) S_s)$$

where N_r is the average number of real plants per real site, and N_s is the number of simulated plants per simulated site.

A very similar calculation may also be used to calculate the parameter for the likelihood of outcrossing between two plants (the pollen dispersal parameter, effectively) by substituting the median seed dispersal distance with the median pollen dispersal distance. In reality, reproduction events occur when pollen from one plant is physically transferred to another. The PopAger and SelectionFinder simulations represent this by synthesising a new genotype comprising alleles from the two individuals chosen as parents, creating a new seed at the same location as one of the parents (see Chapter **3.3.3** for further discussion of the implementation of meiotic recombination used by this project).

When supplied with the estimate of the effective population size, and reasonable values for the median seed and pollen dispersal distances, the number of accessions from the UK represented in the SNP dataset, and an estimate of the mean number of individuals at any given habitable site for the UK population of *A. thaliana*, this equation returns median dispersal distances which, when applied to dispersal of seeds and pollen within a small simulated population, reflect the same rates of dispersal as found in the real-world wild population. Wender, Polisetty and Donohue (Wender et al. 2005) observed a mean seed dispersal distance of roughly 1 metre, with frequent dispersal of 2 metres or more. Since 50% of seeds generated in the simulations were explicitly restricted to the sites at which they were first created (i.e., the same locations as their parents) in order that the simulated population better reflected the 'accession' sub-populations of the wild population (see Chapter **3.3.3** for further discussion), a figure of 2 metres was therefore supplied to the equation shown above. No such figures were available for mean pollen dispersal distance, due to difficulty of observation in light of *A. thaliana*'s low

outcrossing rate (Platt et al. 2010), but outcrossing in *A. thaliana* has long been suspected to be mediated by insects, particularly of the *Syrphidae* (hoverfly) family (Snape & Lawrence 1971). Since flying insects may rapidly disperse pollen over distances greater than that expected from a seed incapable of moving under its own power, a mean pollen dispersal distance of 50 metres was supplied.

Three such sets of parameters were generated: one to best reflect *A. thaliana* gene flow as we know it to occur, as described in the previous paragraph; and two others to reflect greatly higher and greatly lower levels of gene flow in other hypothetical species.

3.3.3 OUTCROSSING AND RECOMBINATION SCALING

The previous chapter discussed how the PopAger (and SelectionFinder) program does not produce a simulated population that accurately represents the volume of gene flow in the much larger wild population until seeds and pollen are set to disperse by a larger distance, scaled to the difference in the number of habitable sites between the simulated and real populations. Similarly, PopAger and SelectionFinder are not capable of representing the volume of meiotic recombination that exists in the real population (due to the volume of gene flow) unless the frequency of outcrossing and crossovers formed per chromosome are increased proportionally to the difference between the number of individuals represented in the simulation and the effective population size of the wild population.

Consequences of the difference in number of individuals between the simulated and real populations become apparent when PopAger and SelectionFinder are supplied with outcross frequency and crossover per chromosome parameters identical to those found in the wild population (derived from (Platt et al. 2010) and (Giraut et al. 2011) respectively): rare haplotypes rapidly go extinct, and the general distribution of haplotype lengths rapidly shifts towards *longer* haplotypes than those observed in the 250K dataset. This can mean only that a

simulation supplied with those parameters fails to represent the level of *population-wide* recombination, or the volume of recombination, responsible for the observed length distribution of haplotypes; a situation that may therefore be rectified by increasing the number of plants in the simulation, the outcrossing frequency, the number of crossovers per chromosome, or a combination of all three**.**

An adequate resolution to the issue of accurately representing the degree of recombination found in the wild population using the unavoidably smaller number of samples available in the 250K dataset is more difficult than the resolution of seed/pollen dispersal distances, especially given that the latter has already been employed. Increasing the number of parameters that are altered from the values observed from the real population increases the dimensionality of the parameter space – simply, allowing more ways to set parameters that cause the simulation to not reflect reality. Since altering the simulated number of individuals per sampling site poses a clear risk of causing the population genetics characteristics of the simulated population to depart from the known characteristics of the wild population (see previous sub-chapter), and since PopAger relies upon those population genetics characteristics to draw conclusions, the number of simulated plants per represented site remained set to the value typical to the wild population. However, the difference in size between the number of simulated individuals and the effective population size of the wild population therefore necessitated large changes to both the outcrossing rate and the number of crossovers per chromosome in order to bring the simulated population's volume of recombination in line with that estimated for the real population, as described below.

As with seed/pollen dispersal distance parameters, altering the outcrossing and recombination parameters risks changing the demographic model that the simulation is based upon to the point that it no longer approximates reality, and

as before, conclusions should therefore not be drawn unless an *a posteriori* test of relevant properties shows that the simulated population approximates the real population. In this case, haplotype length distributions (i.e., the numbers of genomic loci occupied by each haplotype) – see Figure **15** – were used as the validation that outcrossing/recombination was scaled correctly. Data for this validation came from an *a posteriori* analysis of the relative frequencies of length (in number of contiguous genomic loci) of haplotypes that were re-generated after running SelectionFinder using the same parameters (SelectionFinder uses the same methods of population simulation as PopAger – see Chapter **4.3.2**). The distribution of haplotype lengths in the genotypes of this simulated population roughly approximated that reported in the real population in Chapter **2**, though the simulation also generated a small number of haplotypes with lengths greater than any observed in the wild population.

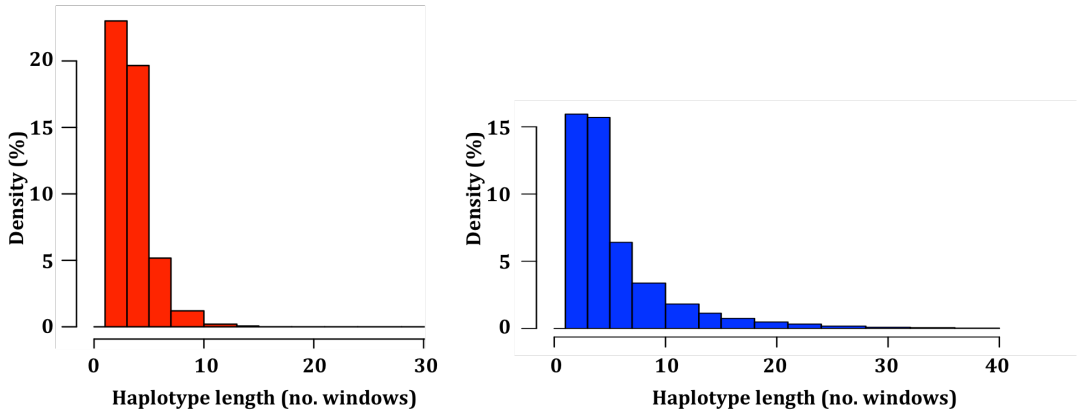


Figure 15 Histograms of haplotype length distributions across wild (**RED** plot) and simulated (**BLUE** plot) populations. Note that the simulated population skews slightly towards longer haplotypes. This is most likely an inevitable consequence of drift, caused by the relatively small size of the simulated population. The consequences of this are discussed in Chapter **3.4.2**.

The potential consequences of altering these two parameters are discussed further in Chapter **3.4.2**).

Initially, this second form of scaling worked around two parameters: population size and outcrossing rate. The discrepancy between the numbers of individuals constituting the wild and simulated populations, along with the

2observed wild outcrossing rate, were first used to estimate an appropriate value for the outcrossing rate in the simulated population, like so:

$$\mathbf{O_s = (P_r / P_s)\, O_r}$$

where O_r and O_s are the outcrossing rates for the real and simulated populations respectively, and P_r and P_s are the number of individual plants in the real and simulated populations respectively.

When supplied with figures reflecting the outcrossing and recombination frequencies observed from the wild population, this equation indicated a required outcrossing frequency parameter of several thousand percent, indicating that even if the simulated outcrossing rate were set to 100%, the volume of recombination in the simulated population would still fall short of that projected to exist in the real population. It was therefore decided that an increase in the number of crossovers per chromosome was necessary. Crossovers per chromosome were incorporated into the equation as shown below:

$$\mathbf{C_s = ((P_r / P_s) * (O_r / O_s)) / C_r}$$

where C_s and C_r are the numbers of crossovers per chromosome in the simulated and real populations, respectively. When the simulation outcrossing frequency parameter is set to a practical value (i.e., representing 100% reproduction via outcrossing or less) and total desired number of simulated plants is supplied, this formula returns a value for the number of crossovers per chromosome to implement in the simulation.

3.3.4 POPAGER: A TOOL FOR DEMOGRAPHIC HISTORY INFERENCE FROM POPULATION STRUCTURE

Results from Chapter **2.3.5** appear to show that, as proposed by Platt *et al.* (2010), the 'isolation by distance'-type population structure commonly found in wild *A. thaliana* is a phenomenon that progressively emerges over time as a

population becomes ever more established. It therefore followed that the rate of emergence of specific population structure could be utilised to estimate the number of generations that have passed since the population in question was initially founded. The PopAger tool described in this chapter was based on this premise. The tool creates an *in silico* population using the genotype data and the demographic parameters calculated in the previous chapter, and published figures and data described at the beginning of this chapter. An overview of PopAger function is presented in Figure **16**.

The PopAger tool works on a relatively simple principle. Starting from a situation in which a set area is uninhabited, it introduces a preset number of simulated individuals. The simulation represents the production of offspring by these founder individuals, which then later constitute the next generation. Each offspring is assigned a genotype as it is generated. Those offspring may be distributed some distance from their parent. Since there are a finite number of growth sites available in this method, only a fraction of the offspring produced by a generation will persist to form the adults of the following generation once the population spreads over the full available range. Upon the selection of offspring to form a new generation (or, depending on the selected setting, after the completion of n generational cycles) the population structure of the simulated dataset is compared with that of the initial dataset gathered from the wild population; and, if the two sets of measurements are sufficiently similar, then the simulation is ended and the number of simulated generations taken to reach the stage is returned as a result. Together, these steps enable a simulated population to propagate from its initial founding to a state similar to the extant wild population. This may be repeated any number of times in order to obtain a statistical range of predictions, and it is recommended that users do so if possible. Each of these series of iterated steps will now be described in greater detail.

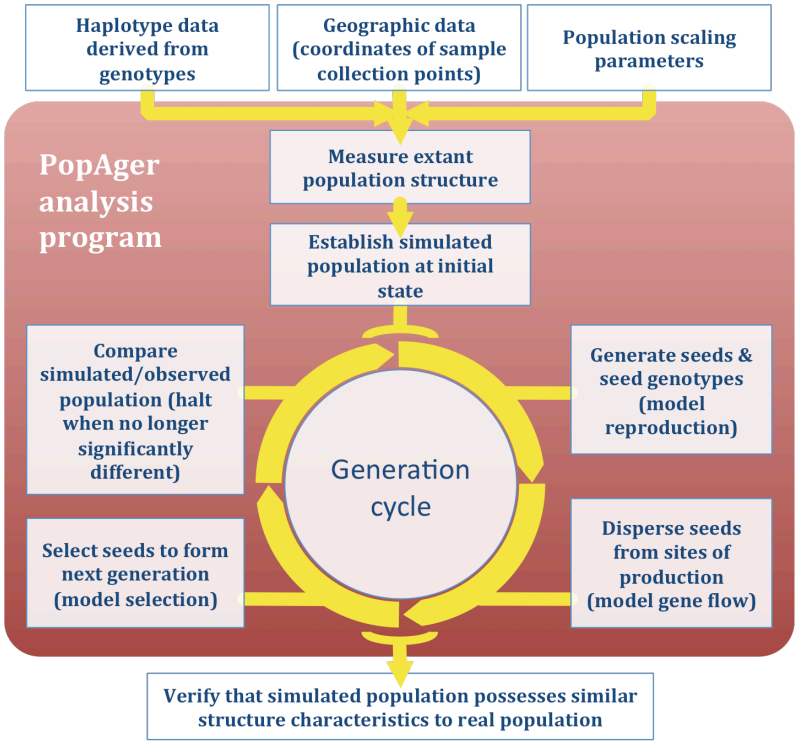


Figure 16 PopAger analysis overview. Flow chart showing the PopAger analysis process. This analysis was designed to estimate the number of generations required for a population within an arbitrarily selected area to develop to its currently observed state under a supplied model of demographics (movement and distribution of individuals), reproductive characteristics and genetic diversity. Initially, measurements of the extant population are taken, in order to act as a baseline for periodic comparison as the simulation progresses. In theory, any given population genetic metrics may be calculated, but the metric used in this project was that of a linear regression of genotypic similarity of samples versus distance between sampling sites (see Chapter 2.[X] for examples of this type of measurement taken from extant *A. thaliana* populations). This metric was selected here both for its relevance to the known 'isolation by distance' population structure of *A. thaliana* (Chapter 2.[X], [Platt 2010]) and for its simplicity of repeated calculation. In the simulation's initial state, all samples within the area of interest (in the context of this project, the UK) are cleared away. A founder individual or small population is then moved from the set of samples taken from sites outside the area of interest and distributed across the listed habitable sites within the area of interest according to the rules of individual movement specified by the demographic model. In the generation cycle, PopAger begins to apply the rules of reproduction specified in the model in order to generate new genotypes, through processes designed to represent either self-fertilisation or outcrossing (see Chapter [X.X for a full description). These genotypes may be moved to habitable sites other than those of their parents, in order to represent the seed dispersal/migration parameters of the supplied model (see Chapter [X.X]). PopAger also limits the number of individuals allowed to exist at any given habitable site, in order to prevent the unrealistic situation of a population experiencing perpetual exponential growth. In this project, [generally neutral, attempt some measure of selection?]. This generation cycle repeats until the general structure observed in the extant population re-emerges in the simulated population. The simulated population may then be examined in greater depth, in order to more rigorously ascertain whether the demographic model accurately describes the real population.

For the sake of practicality, this approach makes several simplifying assumptions, each of which will be discussed in the following 3 paragraphs:

Firstly, that the isolation by distance population structure is determined entirely by distance between sites, rather than by additional geographic and

ecological factors such as physical barriers to gene flow. In practice, the distribution of sampling sites in the dataset means that the majority of gene flow occurs between sites that are located near to each other, so gene flow rarely occurs across regions that are less densely inhabited by *A. thaliana*. A secondary assumption is therefore that the sampling sites represented in the 250K dataset are roughly representative of the distribution of *A. thaliana* across the UK.

Secondly, that the population structure and composition of external populations does not significantly change over the course of the analysis. Since the UK population appears to have been established for a shorter duration than the populations on the European mainland (see Chapter **2.3.5**), it is likely that the mainland populations had already established a stable population structure when the UK population was initially founded.

Thirdly, that all genotypes present in the population are selectively neutral. The majority of genotypic variation is selectively neutral, or nearly so (Kimura & Ota 1973; Ohta 1992). This means that gene flow of neutral alleles is likely to be the most significant factor in the degree of similarity between two accessions (Holderegger et al. 2006), and conversely, that alleles under the influence of selective pressures are responsible for only a small amount of variation in the overall degree of genotypic similarity between any given individuals. Since PopAger compares results based on the overall genotypic similarity of individuals, this means that a simulation modeling selective neutrality is still likely to closely approximate most gene flow occurring in the wild population.

It is, in fact, possible to attempt to improve upon the first of these simplifying assumptions. Pollen and seed dispersal parameters may take the geographic distances separating the sample sites as their basis for the calculation of dispersal likelihood, but these figures for likelihood of dispersal may also be taken from other sources. Clustering results from the program Structure may also be used to give an estimate of this figure between any two samples in the

dataset. This may be achieved by retrieving the result indicating the inferred ancestry of a sample collected at one site to the cluster to which the sample collected at the other site belongs. Using this data may give a subtler means of modeling gene flow via distribution; since it is drawn from measurements of the wild population, this data is likely to reflect the barriers to gene flow present in the real environment that simple straight-line distances fail to capture.

The dataset is first divided into two groups: the 'population' group, comprising all of the samples gathered from within the area under investigation (in the case of this project, the 'population' group would be all samples gathered from the UK); and the 'background' group comprising all other samples (in this project, the 'background' group is the samples gathered from mainland Europe). The genotypes of samples may be read either as haplotypes, or as 'hit' clusters (see Chapter **2**). For speed and scientific quality of processing, it is recommended that 'hit' data is used. Prior to the start of the simulation, a measurement of the structure of the population within the 'population' area is taken by measuring the pairwise dissimilarity of sample genotypes and the geographic distance separating their collection sites, and plotting a linear regression from that data (as described in Chapter **2**).

Simulated plants are allowed to exist only in certain specified locations: those at which samples in the 250K dataset were collected. Since the global coordinates at which each sample was collected are included with the 250K dataset, this allows simulated individuals with appropriate genotypes to be dispersed across a geographic area in a manner reflecting the structure and diversity of the real population. A simulation of discrete individuals in this manner neatly avoids the mathematical complications in attempting to model isolation by distance-type population structure using other approaches attempted in the past, such as Gaussian diffusion (Platt 2012, *pers. comm.*). The locations of these sampling

sites, both within the UK and across the European continent, are shown in Figure **4**.

At the beginning of the simulation, the 'population' area is cleared of all data, and a preset number of seeds – constructed from genotypes of plants in the 'background' area – are caused to disperse into the 'population' area. Genotypes of samples from the 'background' area are maintained in memory, though are not propagated and moved as genotypes in the 'population' area are. They remain available, however, in order to later represent new genotypes appearing in the 'population' area through migration from, or outcrossing with, an external population, should the simulation have been set to incorporate these factors.

Following this founding event, the initial generation of simulated plants within the 'population' area is caused to produce seeds of its own, and to disperse them according to parameters set out in Chapter **3.4.1**. Each plant produces a preset number of seeds; in order to allow the population to increase its numbers, this value should be set to >1. Typically a value of 2 was used in order to conserve computational resources while maintaining the population's ability to expand; this parameter may be set higher, and indeed it is recommended to do so if possible; however, production of two seeds per plant consumed almost all available computations resources when using high-density haplotype data. When assembling the genotype of the seed, two options are given: to produce a new genotype via self-fertilisation, or via outcrossing. The outcrossing mode of reproduction was set to occur at a low frequency of 3%, in accordance with published observations (Platt et al. 2010). Mode of outcrossing was selected prior to each individual reproduction event, using an appropriately weighted random choice.

Simply copying the genotype of the parent represents construction of genotypes via self-fertilisation events; this reflects the natural near-complete

lack of heterozygosity observed in wild *A. thaliana*. The construction of genotypes via outcrossing events is somewhat more complex, however.

Outcrossing events are represented through a series of processes. Firstly, a second parent is selected. This second parent may reside at any of the sites represented in the dataset, though a weighted random choice is utilised in order to proportionally favour the selection of parents located geographically near to the first parent over geographically more distant plants. The weights assigned to other plants in this selection are calculated with an exponential decay function, the exponent of which is given by the 'pollen dispersal parameter' described in Chapter **3.3.2**. Selection of parents to participate in outcrossing events in this manner represents the dispersal of pollen in the wild population.

Secondly, the genotypes of the two parent plants are retrieved and subjected to a process intended to reflect the recombination of genotypes through meiotic crossover. A number of sites at which recombination will occur are selected along the length of each chromosome. This number is set manually, and was normally set to 2, in accordance with measurements of crossover frequency reported in the literature. Additionally, variation in the number of crossovers set to occur on any given chromosome has been added; this variation is set to follow a normal distribution around the supplied value for the mean number of crossovers per chromosome. Its extent is determined by a second figure – the standard deviation of that variation (a figure typically set to 1). This means that the pattern of recombination in the simulated population better approximates that reported in the literature. There is facility for crossover rates to be favoured at some locations over others, via another weighted random choice, though its usage was not deemed necessary within the context of this project. In order to produce a recombinant genotype, one of the two parents is first chosen at random. All of the haplotypes belonging to that parent between the start of the chromosome and the first crossover point are then assigned to the new

genotype; following that, the haplotypes belonging to the other parent between the first and second crossover points are also added to the new genotype. This alternation between parental contributions continues until a complete chromosome is assembled, possessing parts of the genotypes of both parents.

It should be noted that this is not an entirely accurate reflection of the meiotic process; in reality, crossovers between P-generation chromosomes only occur in F_2-generation offspring, rather than F_1-generation offspring as described here, since an individual organism's genotype is represented in this dataset in a state of haploidy. However, and since outcrossing is a relatively uncommon phenomenon in *A. thaliana* (leading to the low observed rate of homozygosity in the wild population), this representation of meiotic recombination provides a sufficiently close approximation of the results of the process.

The process of recombination is represented in essentially the same way whether haplotype or 'hit' data is used, though use of haplotype data requires a slightly more complex solution on account of haplotypes spanning multiple 'windows'. It is likely that a crossover point will fall within the bounds of a haplotype, bisecting it; in this case, the descendent genotype will receive only a fraction of the haplotype that exists in the parental genotype. This reflects the gradual breakdown of haplotypes as they age, as predicted by Kimura & Ohta (Kimura & Ota 1973). This is implemented in PopAger through a mathematical tool known as an interval tree. The time required to process this problem is significantly greater than the time required to generate new recombinatory genotypes when each window is treated essentially as an independent locus, as is the case when 'hit' data is used. Additionally, haplotype data may leave gaps within the genotype of some individual organisms, in which no haplotype was conclusively assigned. As crossovers shorten haplotypes, the number and size of gaps may begin to grow, meaning that data is lost and the power of the analysis is reduced. Again, this is not an issue when 'hit' data is used, since genotypes have no gaps when stored in this format.

Following the creation of seeds and their respective genotypes, the seeds are caused to scatter from the sites at which their parent individuals reside, in order to represent seed dispersal in the wild population. Dispersal of seeds is handled in a similar manner to the selection of parents for outcrossing events; the distances from a seed's current location to each of the other sites within the 'population' area are retrieved, and are used to assign an appropriate weight of likelihood to that particular migration path, in preparation for a weighted random choice of seed destination. As with pollen dispersal, seed dispersal weight is calculated with an exponential decay function using the seed dispersal parameter calculated in Chapter **3.3.2** as the exponent.

As later results will show, the isolation by distance population structure can be shown to emerge as a consequence simply of these outcrossing and seed dispersal characteristics. Additional characteristics were also added for the purposes of this analysis, however, in order to reflect additional knowledge of *A. thaliana* and its wild habits.

In the wild, *A. thaliana* tends to form small groups of genetically identical (or near-identical) individuals that, due to the short dispersal distances of seeds (Wender et al. 2005), possess genotypes that may survive relatively intact in the same location for many generations (Bomblies et al. 2010). A simulation in which seeds more often disperse to a different site than not, then, reflects this reality poorly. PopAger attempts to rectify this discrepancy by making each plant in the parental generation create a second set of seeds that are not dispersed and do not outcross (so are therefore genotypically identical to their parents), but instead remain in the same site as their predecessors. This has the dual effect of continuing to maintain a representation of a local population even if no other seeds are dispersed there in any given generation, and also of more adequately representing the likelihood of a group of plants at a site being entirely replaced. Since the non-mobile seeds are all produced in a manner that replicates self-fertilisation, PopAger increases the frequency of outcrossing in

the production of mobile seeds in order to maintain the correct frequency of seeds produced through outcrossing events throughout the complete set of seeds.

Additionally, PopAger is capable of simulating migration of seeds from external populations (*i.e.,* those represented by samples in the 'background' area). This is achieved by first selecting a set of samples from the 'background' set that are to distribute seeds into the 'population' area (using a weighted random choice based on distance from the 'population' area), and then by dispersing seeds possessing the genotypes of the selected samples from their points of origin into the 'population' area in precisely the same manner as already discussed. As with the introduction of stationary seeds, this feature may be disabled as per the requirements of an analysis.

In order to prevent the population simply increasing in number exponentially to the point at which new generations become impractical to process (a situation which, in any case, does not reflect real populations beyond their initial founding and expansion), the number of seeds allowed to develop into the next generation's adult plants was restricted by enforcing an upper limit – controllable by the user – on the number of seeds at each site that could be selected to do so. Should the number of seeds at any given location exceed that threshold, PopAger simply selects seeds at that site at random until the maximum number permitted to enter the next generation is reached. Note that if a model called for seeds possessing certain genotypes to be more likely to appear in the next generation than others at this particular site (*i.e.,* if the model required an implementation of natural selection) this may be achieved by using a weighted random choice, with appropriate weights for each seed based on its genotype, in place of a purely random choice. Since neutrality was taken as a simplifying assumption for the purposes of this analysis, and since a comprehensive modeling of selective pressures across all sites requires more data than is currently available in any case, no selective pressures were

modeled in this analysis. Note also that the value supplied also acts as an effective and simple means of controlling the extent of genetic drift within the simulated population, but that processing time and memory usage increase linearly with the number of plants allowed at each site.

Once the seeds that are to form the next generation of plants have been chosen, the structure of the simulated population within the 'population' area may be measured. Depending on the setting supplied by the user, this measurement may be taken every generation, or every *n* generations. As with the initial measurement of population structure, the measurement of population structure within the simulated population is achieved by tallying similarity of genotypes between pairs of plants versus the distance separating them, and fitting a regression to that data. The regression measured from the current simulated generation is compared against the regression measured from the wild population using an analysis of covariance (ANCOVA), testing the hypothesis that the two distributions are the same. Should the ANCOVA return a p-value of >0.5 (meaning that sufficient evidence that the two distributions are different cannot be found), the simulation ends and the number of generations since founding is returned as a result. The simulation also ends once a pre-set number of generations have passed, in order to prevent overrun.

The measurements of genomic similarity vs. distance in both the wild and simulated population may be plotted at each measurement interval if requested by the user.

3.3.5 RESULTS FROM POPAGER

The PopAger tool was, initially, used to demonstrate that the emergence of isolation by distance-type population structure reported in Chapter **2.3.5** is a consequence of outcrossing (and thus meiotic recombination) and short-range dispersal across a broad geographic range. The 'population' area was set to include all sample collection sites from the UK.

With the exception of the first experiment described immediately below, genotypes for the simulated plants were read in from the 'hit' dataset generated as part of the initial survey of haplotypes described in Chapter **2.3.2**. 'Hit' data was used in this analysis, as in Chapter **2**, in order to facilitate a fairer comparison with prior work by Platt *et al.* (2010), and also with earlier findings reported in Chapter **2.3.2**.

In the first experiment carried out with PopAger, a single founder was selected at random from the 'background' genotypes on the European mainland, and introduced to the 'population' area (the British Isles). No further genotypes were introduced in subsequent generations. This experiment was run for 200 generations.

Genotype data was supplied in the form of haplotypes for this experiment, in order that the effect of recombination was made clear through the implicit tracking of the fate of only a single founding genotype. Specifically, this analysis utilised haplotype data produced by the second version of the program developed to find haplotypes (see Chapter **2.3.3**). Haplotype data produced by this method contained gaps – windows in which no haplotype was specified.

Due to the gaps in the set of haplotypes assigned to the founding individual – which were represented within PopAger as loci with no explicitly stated genotype, but at which crossovers could still occur – it was possible to observe the original haplotypes shortening over time, as described in Chapter **4.1.2**, due to recombination. In a real population, the gaps in the genotype data would contain an alternative pattern of alleles that would replace the alleles of the haplotype when recombination occurs, but since the alternative genotypes were rendered effectively invisible in this analysis, direct observation of the dispersal of the genotype across the UK and the increasing dissimilarity of the founder's descendants was possible.

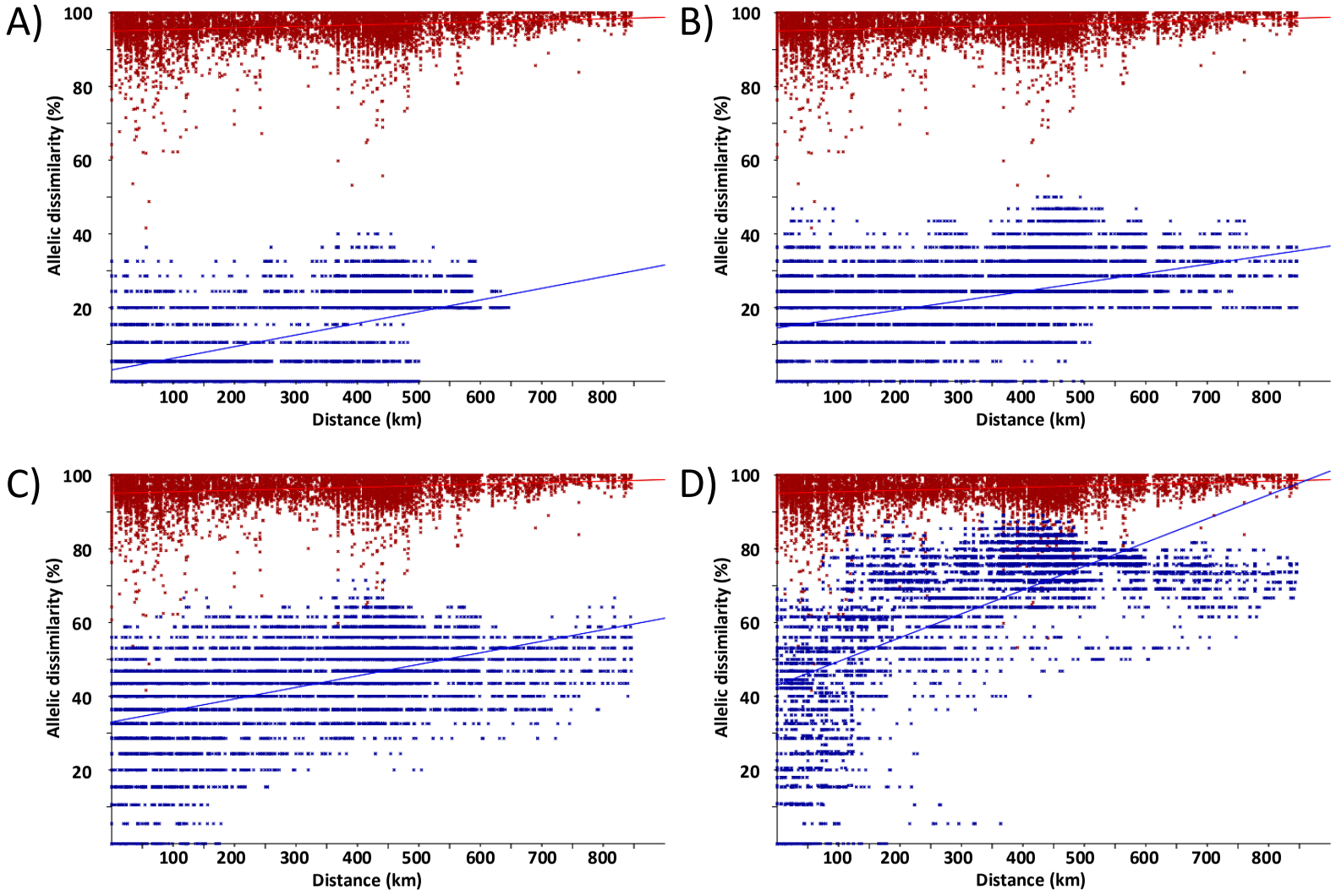

A)
B)
C)
D)
Allelic dissimilarity (%)
Distance (km)
100
80
60
40
20
100
200
300
400
500
600
700
800

Figure 17 Initial demonstration and verification of PopAger action in which a population spanning the UK was caused to arise from a single founder. Measurements of genetic dissimilarity vs. separating distance between individual plants are shown at generation 10 (**a**), 30 (**b**), 75 (**c**), and 200 (**d**) simulated generations, for both the wild population (**RED** points) and the simulated population (**BLUE** points) at that time-point. The simulation from which this time series was generated was slightly modified in order to best demonstrate the effects of seed dispersal and outcrossing as prescribed by the demographic model parameters generated in Chapters **3.3.2** and **3.3.3**, via the simulation of the arrival and establishment of a novel genotype into the UK. As expected, the descendant generations migrate to other habitable sites, and outcrossing and subsequent recombination cause their genotypes to diverge from that of the founder and from each other. The pattern of isolation by distance soon becomes apparent, and becomes progressively more similar to that observed from the wild population.

As the population spread across the sites represented in the simulation, outcrossing began to occur between plants at different sites, causing haplotypes to gradually reduce in length and eventually become eliminated from some individuals. Figure **17** tracks the progression of this differentiation over 200 simulated generations, showing both a clear pattern of isolation by distance, and a steady emergence of increasing differentiation between plants at different sampling sites. As the simulated population expanded to fill all available habitation sites, recombination began to increase the dissimilarity between individuals; due to the gaps inherent to the haplotype dataset – windows within the genome at which no haplotype is recorded – haplotypes present in the founder's genotype were occasionally lost through recombination. As meiotic recombination caused haplotypes to progressively shrink and eventually go extinct from the population (see Chapter **4.1.2**), the analysis interpreted this as a decrease in the level of genetic similarity between samples. Examined across an entire population, this individually haphazard loss of haplotypes and decrease of similarity was observed to produce a steady decrease in similarity between accessions, and an emergence of isolation by distance-type population structure, progressively approaching the observed distribution of similarity vs. separating distance.

Following this initial demonstration, continual immigration into the UK was enabled, based – as with dispersal within the UK – on distance between sampling sites. Three simulations with different dispersal parameters were

prepared, in order to investigate the potential utility of this approach to species possessing parameters different to those of *A. thaliana* – for instance, more widely dispersing or highly invasive species.

- Set 1 was created to reflect a species with very short median pollen and seed dispersal distances, reflecting a species well adapted to a specific and stable habitat, with little requirement to move elsewhere. Unscaled median dispersal distances were set to 0.1m for seeds and 1m for pollen.
- Set 2 was created to reflect the wild state of *A. thaliana* as best possible, given published data and knowledge. Unscaled median dispersal distances were set to 2m for seeds and 50m for pollen.
- Set 3 was created to reflect a very invasive species, with high dispersal parameters reflecting a ready spread over relatively long distances – for example, through dispersal via animal vectors. Unscaled median dispersal distances were set to 20m for seeds and 1000m for pollen.

Parameter set 1 was run for a total of 300 generations. While the simulated population rapidly began to show roughly similar levels of genetic similarity to that observed in the wild population, the actual geographic extent of the simulated population expanded only slowly – and, by the 300th generation, still covered only a fraction of the available range, with little sign of advancing further.

Parameter set 2 completed its run after just ten simulated generations, with the resulting population precisely matching the observed wild population. An ANCOVA between the wild and simulated populations showed no significant difference between the populations. Regression analysis of each population also showed both populations to follow the same trend.

Parameter set 3 was also run for a total of 300 generations. The simulated population rapidly attained a degree of genetic diversity greater than the observed population. Regression analysis also showed that this population did

not follow the isolation by distance model present in the wild population; degree of genetic similarity between individuals remained, on average, consistent across all distances.

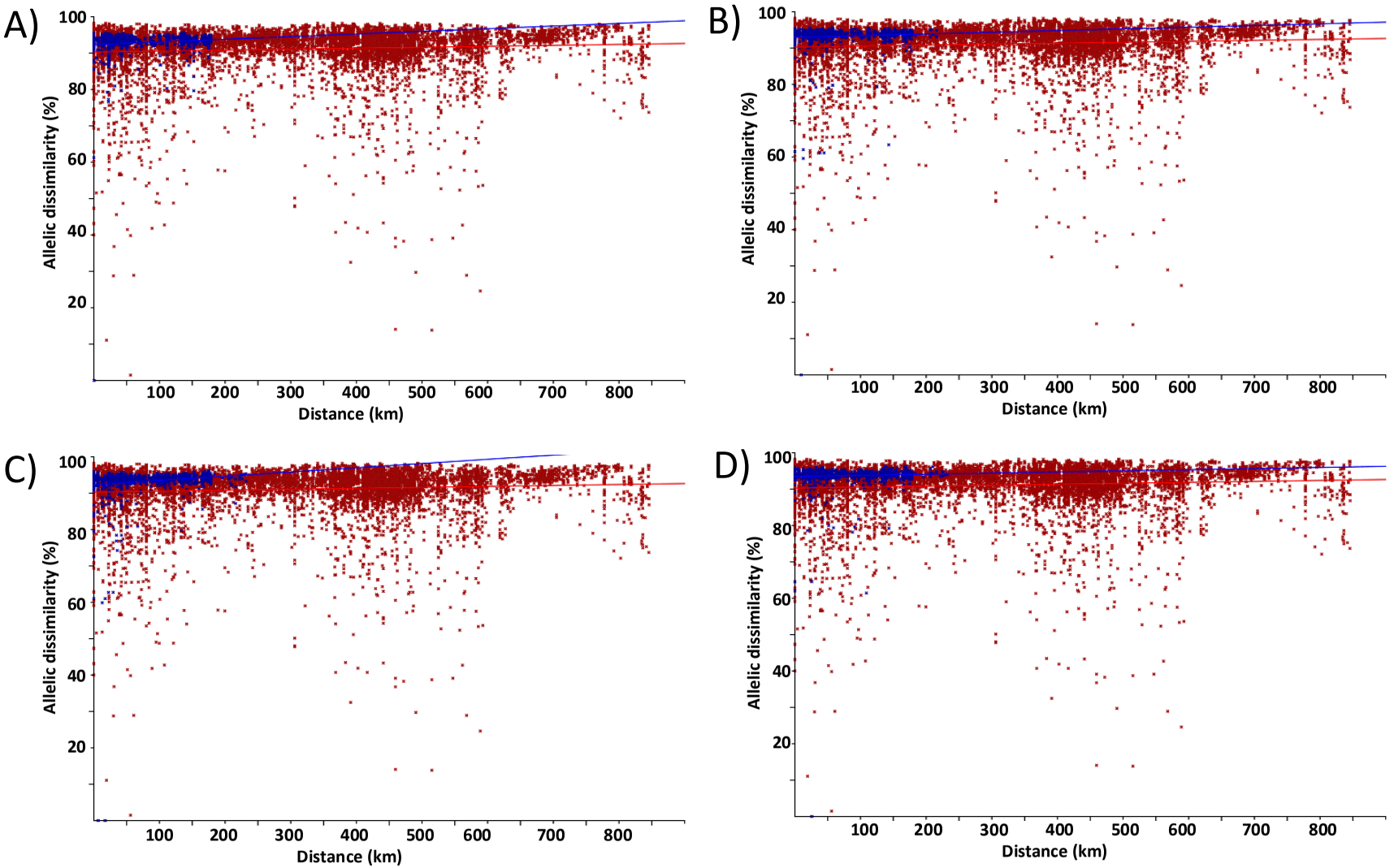
Time series plots showing the emergence of structure within the simulated populations under parameter sets 1, 2 and 3 are shown in Figures **18**, **19** and **20** respectively.

3.3.6 VERIFICATION OF POPULATION STRUCTURE MODEL

In order to demonstrate that the demographic model is sufficient to explain the emergence of observed population structure, the principal coordinate and Structure clustering analyses were repeated on the populations produced during simulations.

Upon the completion of the PopAger run using parameter set 2, the genotypes of simulated plants across the British Isles were re-integrated with those taken from the 250K dataset. PCA clustering of genotypes produced by PopAger (Figure **21**) shows the same pattern of clusters as that produced by applying the same analysis to the genotypes of the wild population (see Chapter **3.3.1** and Figure **10**). Re-clustering the simulated genotypes with Structure (Figure **12**) shows that the frequencies of genotype clusters in the simulated population did not become established at exactly the same frequencies as observed from the wild UK population; however, this is likely to be at least partially attributable to genetic drift, given the small size of the simulated population. Further, Structure analysis shows that the simulated population also exhibits a degree of admixture comparable to that observed from the wild population.

Figure 18 (Next page) Application of PopAger to a hypothetical low-dispersal species showing the extent of population structure after 50 (**A**), 100 (**B**), 200 (**C**) and 300 (**D**) generations. The measurements of dissimilarity/distance taken from the simulated genotypes (**BLUE** series) are plotted against the same measurements taken from the extant *A. thaliana* population (**RED** series) for comparison. Note the inability of the population to expand beyond a small subset of available sites.


A)
B)
C)
D)
Allelic dissimilarity (%)
Distance (km)
100
80
60
40
20
100
200
300
400
500
600
700
800

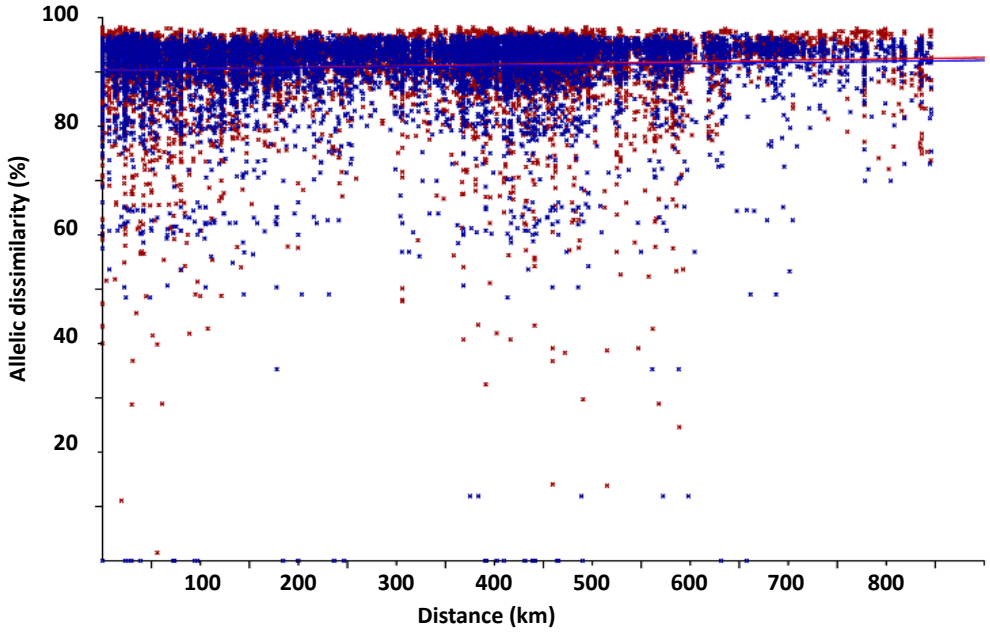
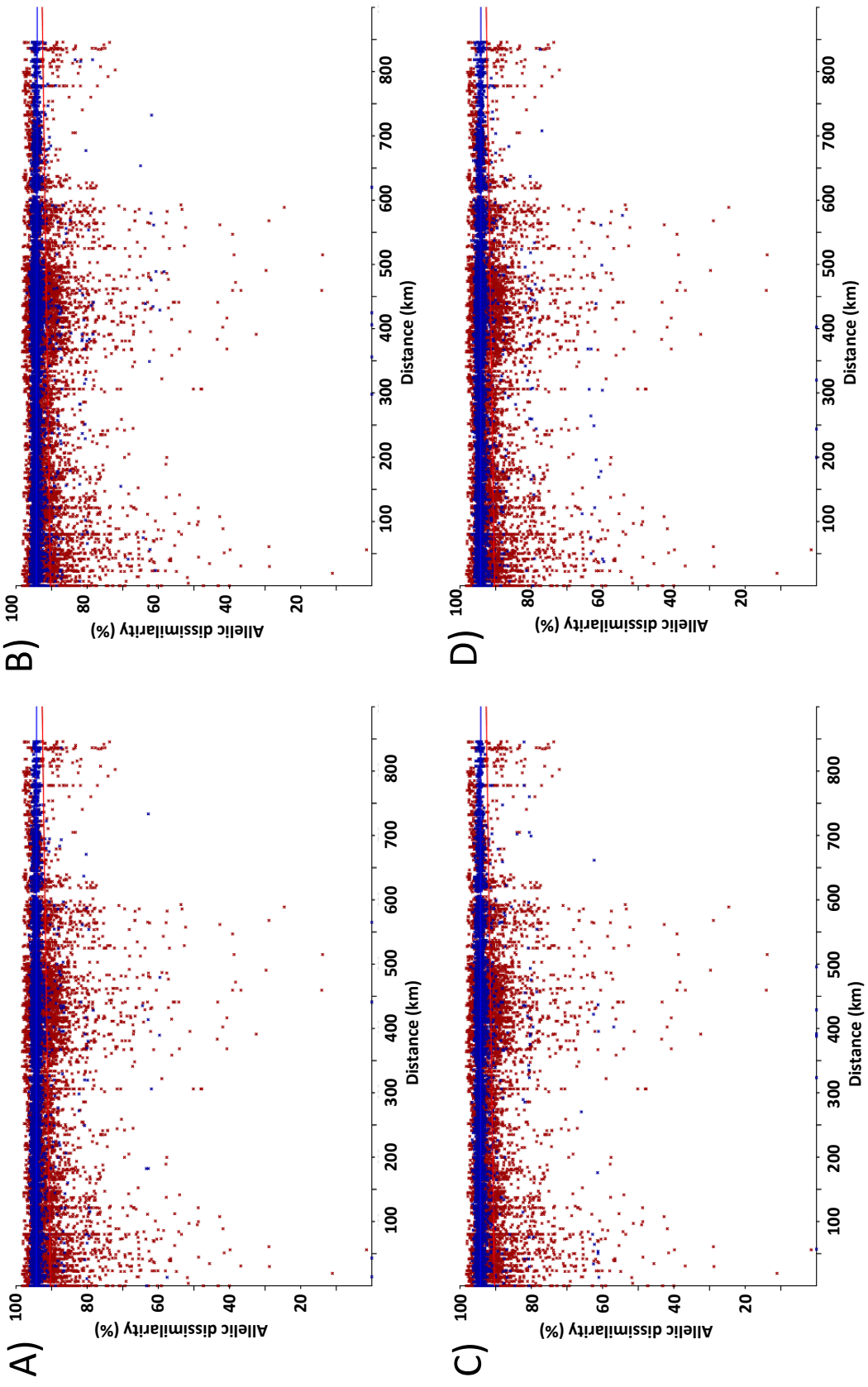


Figure 19 Application of PopAger to simulation following best estimates of *A. thaliana* parameters. This simulation was run under parameters reflecting those derived from measurements of the extant population as sampled in the 250K dataset, from published figures on frequency of recombination and outcrossing (see Chapters **3.3.2** and **3.3.3** respectively), and from reasonable assumptions of typical *A. thaliana* seed and pollen dispersal parameters. Data collected from the simulated population is shown in the **BLUE** series. Data from the extant population is shown in the **RED** series. Despite the possibility of introducing a large degree of error through substantial parameter scaling (see Chapter **3.4.2**) and unaccounted-for stochastic effects resulting from the relatively small number of individuals in comparison to the extant population, the simulation produced a set of genotypes showing no significant difference from the extant population in the general distribution of genotypic similarity across geographic distances within 10 simulated generations. This indicates that while the PopAger tool may not be helpful in the context of inferring the duration of a species' occupation of an area, the demographic model employed in this simulation facilitates the accurate representation of a large wild population *in silico*.

Figure 20 (Next page) Application of PopAger to a hypothetical high-dispersal (invasive) species showing the extent of population structure after 50 (**A**), 100 (**B**), 200 (**C**) and 300 (**D**) generations. Parameters for mean seed and pollen dispersal were set in order to represent a species which is more readily dispersed over long distances than *A. thaliana*. As in Figures **17** and **18**, the **BLUE** series shows measurements from the simulated population, and the **RED** series shows measurements from the extant *A. thaliana* population. The distribution of genotypic similarity over distance in the simulated population follows a trend (shown by the blue regression line) much more closely than that of the extant *A. thaliana* population. That trend is essentially flat (*i.e.*, isolation by distance is essentially non-existent). A lack of isolation by distance structure is an expected outcome in species capable of frequent dispersal over long distances, since such species may be expected to establish an unstructured population more closely approximating Hardy-Weinberg equilibrium. See Chapter **3.3.5** for a full discussion of this figure as well as Figures **17** and **18**.

Figure 20


A)
B)
C)
D)
Allelic dissimilarity (%)
Distance (km)
20
40
60
80
100
100
200
300
400
500
600
700
800

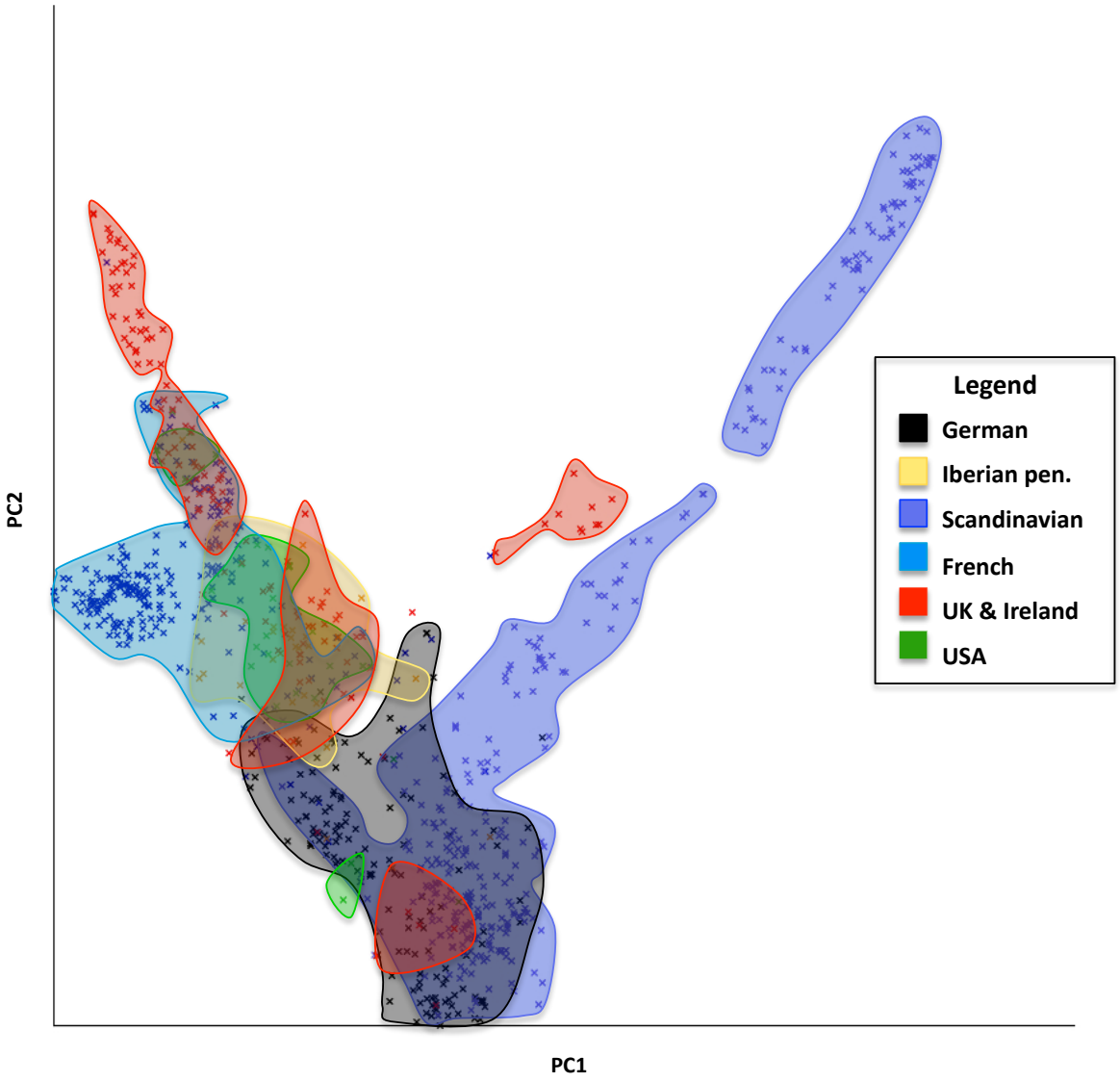


Figure 21 Principal Coordinate Analysis (PCA) of haplotypes in a global sample of *Arabidopsis thaliana*, including UK genotypes established at the conclusion of a PopAger simulation. Genotype data from plants simulated in PopAger run reflecting best demographic estimates at sampling sites across the British Isles was re-integrated into the full 250K dataset and subjected to a repeat of the principal coordinate analysis as described in Chapter **3.3.1**. The pattern of UK genotype clusters returned by this analysis is highly similar to that returned when the analysis is applied to the 250K dataset (Figure [**X**], indicating that the demographic model adequately explains the observed general distribution of genotypes across the UK.

3.4 DISCUSSION

3.4.1 ECOLOGICAL ANALYSIS OF GENOMIC DATA

Principal coordinate analysis of the entire set of haplotype genotypes shows that individuals within the UK are often more genetically similar to populations from parts of the mainland than they are to other individuals within the UK. Clustering of the principal coordinate data shows that there are five distinct genotypic groups in the UK - an observation supported by the Structure analysis also carried out in Chapter **3.3.1**. Since work in the previous chapter concluded that the UK population was founded more recently than mainland populations, similarity between UK and European genotypes therefore most likely indicates migrations from the mainland to the UK, rather than in the opposite direction.

Initially, it was suspected that one of the clusters derived from the principal coordinate analysis – specifically, the 'UK-only' cluster – represented either a much earlier migration event than the other apparent migration events responsible for the constitution of the current UK population, followed by differentiation of the UK founding population during a subsequent period of relative isolation; or a source of genetic variation in the mainland population that the 250K dataset had failed to capture. However, the PCA verification step described in Chapter **3.3.6** shows that the genotypic variation present in the mainland population, together with the demographic model presented, are sufficient to explain essentially all observed structure in the UK population.

These results are broadly consistent with those of Horton *et al.* (Horton et al. 2012), who examined the structure of the European and North American populations using the same 250K SNP data as utilised in this project. Through a principal component analysis of SNP alleles, they found a distinct pattern of differentiation between various groups of accessions in mainland Europe, including a general differentiation between most UK and mainland European accessions, with the exception of a small number of accessions sampled from

the UK that clustered amongst German/French, Spanish and Scandinavian populations. Crucially, while Horton *et al.* (2012) demonstrated the general structure of clusters within the data, this paper does not specify which accessions cluster differently to the rest of the UK population, nor does it reveal the exact differences in the genetic constitution of the accessions that cluster so.

Analysis of the distribution of UK clusters across habitat types shows significant deviation from the null hypothesis only in samples clustering most closely with Scandinavian accessions. This may suggest that those Scandinavian-type genotypes are adapted to that habitat. However, this is a very small sample, and this should be taken as no more than a hypothesis for future research to test. The spatial distribution of these clusters across the UK and the continent appears to reflect the conclusion from the F_{ST} analysis that admixture between habitat types is extensive.

The effective population size of the UK population was also estimated, and used to calibrate PopAger model parameters for dispersal. Due to the simple model of dispersal likelihood based on exponential decay as distance increased, relatively small changes in the value predicted for median seed dispersal created large effects on the maximum distance seed or pollen might disperse. As demonstrated, this means that the approach may be applied to populations of species restricted to small clusters of sites with little gene flow between them. However, since an effective framework to deal with these circumstances is already well established, the approach is unlikely to reveal any new information in this context. The strength of the method, instead, lies in the ability to inform on the opposite situation: that of a widespread population with extensive gene flow, as exists in *A. thaliana*, as shown by the measurements of F_{ST} derived from the 250K dataset (Table **4**).

Given the degree to which human beings influence the environment across much of Europe, it is reasonable to expect a substantial degree of human influence in the dispersal of *A. thaliana*. Humans may inadvertently disperse *A.*

thaliana seeds in a number of ways, such as along passage along roads, shipping lanes and railways; through the collection and usage of compost and peat or through the reuse of stones and mortar in the construction of walls. The high-quality genomic data available to this project provides an excellent opportunity to attempt to identify the places in the UK between which this human-mediated dispersal is likely to be greatest. Data stored in Appendix **1** identifies paired sites between which this type of human influence is most likely to occur. These matched pairs could easily be taken as hypotheses to be tested through resampling/resequencing experiments.

3.4.2 SIMULATION PARAMETER SCALING

In the PopAger (and also SelectionFinder) simulations, it was necessary to scale the seed/pollen dispersal distances and recombination/outcross rates of the simulated plants, in order that the small population established within the simulation could be used to model the degree of gene flow occurring across geographic areas in the much larger wild population. This was attempted by altering the scaling factor of the Pareto distribution governing the likelihood of seed/pollen dispersal, and by multiplying the mean outcross and crossover frequencies to match the values predicted to occur per generation across the entire wild population.

This risks causing the simulation to fail to accurately represent the wild population, and therefore risks compromising the usefulness of any analysis based on this method; however, given the data and compute resources available, no better alternative was possible. In some cases, this attempt to represent a large population through a significantly smaller, less widely distributed one has indeed limited the usefulness of the method – the PopAger analysis itself proved unable to represent the progressive establishment of population structure over hundreds or thousands of generations (see Chapter **3.4.3** for discussion of the probable causes of this failure).

However, in other cases, the scaled simulation did prove suitable for recreating aspects of the wild population. The frequency distribution of haplotypes of various length classes in the simulated population approximated that observed in the wild population, though skewed slightly towards greater frequencies of longer haplotypes. This skew may have been caused by a failure to precisely scale the crossover frequency in proportion with the outcrossing rate, leading to, on average, slightly fewer crossovers per outcross, and therefore to slightly longer haplotypes. Nonetheless, the extent of isolation by distance population structure observed in the UK was recreated almost exactly (see Figure **21** and Chapter **3.3.6**), indicating that the simulation was capable of an accurate representation of gene flow in spatial terms, if not in temporal terms. This means that analyses depending on an accurate representation of these characteristics of a population – as SelectionFinder does – may be used to reliably draw conclusions regarding the wild population.

3.4.3 POPAGER ANALYSIS

It was hoped that a simulation of the population given accurately scaled parameters would reveal the amount of time required for an observed structure to emerge. In reality, this did not prove to be the case. A much larger-scale simulation, in which minimal or no scaling of parameters is necessary, may be required to bring that approach to fruition.

Nonetheless, this approach provides useful information. The high degree of accuracy with which the observed population structure was recreated from scaled parameters shows that these parameters most likely reflect the real dispersal parameters very well. Future studies utilising any similar approach may find this information useful.

Under the simple model used in this project, which made no explicit provision for barriers to gene flow besides large distances between habitable sites, and in which parameters were scaled in order that the gene flow and admixture of the whole wild population were represented by a much smaller number of

individuals across a more limited number of habitation sites, cross-channel gene flow was not an uncommon event. Immigration from mainland populations occurred sufficiently frequently to establish a population with the same degree of isolation by distance structure as the wild population in less than 10 generations.

Establishment of the population as it exists today in such a short duration is clearly not representative of the real situation; however, this result does indicate that the population structure observed in Chapter **2.3.5** can be explained predominantly by a dispersal pattern of seed and pollen based on the exponential decrease of the likelihood of transfer between given sites as the distance between those sites increases. Crucially, this means that the method of simulation discussed in this chapter is also directly applicable to the development and application of the SelectionFinder tool described in the following chapter (Chapter **4**).

Under this model, *A. thaliana* may readily disperse genotypes across a large distance over land through a series of shorter dispersals between nearby habitable sites over many generations, but must cross barriers like the English Channel in a single dispersal event. Since the dispersal parameters in PopAger were scaled in order to allow longer dispersals, this 'stepping stone' effect would not have occurred, and long-distance dispersal via land routes would have occurred very rapidly in the simulation. This would explain why the structure observed in the wild population was recreated after so few simulated generations. This explanation has precedent fro theoretical predictions of gene flow and differentiation within an isolation by distance-type population structure made by Kimura and Weiss (Kimura & Weiss 1964).

Therefore, future investigations following this line of reasoning should model as many habitable sites as possible. A future analysis making use of the PopAger method could divide the area under investigation into a grid, and standard ecological survey methods could be used to estimate the number of habitable

sites for a given species within each grid square. Within each square, a randomly positioned set of habitable sites could be supplied to PopAger. Ideally, the overall number of habitable sites should closely match that estimated from the effective population size and average number of individuals per stand (as in Chapter **3.3.2**), in order that the gradual stepping-stone advance can be represented. Since such a simulation would require the storage, access and modification of a substantially larger number of simulated genotypes than used in this project, genotype data used in a simulation of this proposed scale should be less dense than the 250K dataset – and should, in fact, be more similar to the low-density data collected by Platt *et al*. (Platt et al. 2010)

CHAPTER 4: EVIDENCE OF SELECTION FROM GENOMIC DATA

4.1 INTRODUCTION

4.1.1 PLAN OF ATTACK

Work in this chapter aimed to test hypotheses **H4**: Local groups of *Arabidopsis thaliana* within a particular habitat possess alleles at a frequency significantly different from that expected under a selectively neutral model, indicating that selection is acting upon those alleles, and **H5**: Signatures of selection indicating that selection is acting upon alleles associated with disease resistance are shared consistently across populations in different habitat types. Taken, together, the testing of hypotheses **H4** and **H5** amounted to a set of tests designed to identify local adaptation of *A. thaliana* subpopulations to unique environmental conditions of habitat type and disease prevalence.

The premise of the analysis central to this chapter was that haplotypes driven to their current frequency by selection – in particular, by a selective sweep – can be identified by the pattern of genetic variation in the population; and, moreover, that the agent of selection may be putatively identified by associating the identified loci with candidate genes previously described by functional experiments in real organisms. The objective was to create a model of a selectively neutral population, which simulates the level and pattern of genetic structure observed in the real population. Theory predicts that haplotypes possessing alleles favoured by selection will not conform to the genetic trends described by haplotypes in the population under the neutral model. The approach is similar to a typical 'case/control' GWAS: while the 'case' dataset is comprised of genotypes assayed from an extant, wild population, a 'control' dataset is derived from a simulated rather than a real population.

Alleles thus identified were then compared with published data describing gene families associated with responses induced by biotic and abiotic environment,

providing an indirect test of Hypothesis **H5**. Mapping results from an experiment carried out to identify loci conferring resistance to a common obligate parasite in wild crucifer populations was also used to provide a more direct test of the hypothesis.

4.1.2 CHARACTERISTICS OF HAPLOTYPES UNDER SELECTION

As has previously been discussed, attributing observed linkage disequilibrium to one cause over another (*e.g.,* to selection rather than to population structure or genetic drift) has historically been a considerable challenge. On their own, measurements of linkage disequilibrium cannot be conclusively attributed to any one phenomenon. Supporting information is required to resolve the situation, such as measurements of the structure of the population from additional, unlinked loci, and estimates of the effective population size in order to estimate the degree of drift. Given this information, it becomes possible to compare the observed linkage disequilibrium with that expected from any of these factors, and thus to determine whether that the observed LD is similar to that expected by, for instance, genetic drift, and dissimilar to that expected from other potential causes.

Seminal work by Kimura & Ohta (Kimura & Ota 1973) shows that under conditions of neutrality, new alleles that arise by mutation within a population follow a predictable pattern of frequency due to drift. A new allele will emerge, may tend to drift towards fixation for a time, and will eventually wane and disappear altogether from the population. The amount of time over which this typically occurs, and the maximum frequency an allele might be expected to reach (identified as θ), are determined in a neutral unstructured population entirely by the effective population size, which controls the degree to which genetic drift affects the allele frequencies from one generation to the next. This model can, of course, be expanded to encompass a structured population, and to encompass haplotypes spanning multiple alleles.

Since a haplotype spans a number of loci, its genetic length would be expected to alter its expected value of θ. A longer haplotype is more likely to have a crossover fall within its length during meiosis than a smaller one, reducing its length. Consequently, as described in Chapter **2.1**, long haplotypes will inevitably have arisen relatively recently. The expected value of θ for more recently arisen alleles is likely to be smaller than for older alleles, since the chances of a recently arisen allele rising to high frequency due to drift are much lower than for an older allele. Also, should a long haplotype reach a high frequency in the population, it will be subjected more often to recombination, and its average length within the population would decrease. Therefore, values of θ for long haplotypes would be expected to be smaller than values of θ for short haplotypes.

Recombinatory breakdown carries another consequence of relevance to any search for haplotypes under selection. If a haplotype possess an allele that is favoured by selection, then that allele will tend to rise in frequency and move towards fixation, and alleles at nearby loci will also tend to be drawn along to fixation due to the phenomenon of genetic hitchhiking (Smith & Haigh 1974). Under a neutral model, meiotic recombination would be expected to break up a haplotype essentially at random. However, when an allele is under selection, the move towards fixation, and the linkage between adjacent alleles that causes the genetic hitchhiking effect, will cause the part of the original haplotype surrounding the allele favoured by selection to tend to endure in the population for longer than the rest of the haplotype. In other words, the remnants of the haplotype will tend to centre upon the allele upon which selection acts, and the haplotype will be best conserved (and thus most frequent) within the population at and immediately surrounding the allele under selection, breaking down as normal at more distant and less tightly-linked loci. The converse of this situation is that if a haplotype shows signatures of being under selection, the actual loci (at least in terms of genes potentially associated with an adaptive phenotype) upon which selection is acting to drive an allele towards fixation

may be pinpointed by looking for the genomic region in which their haplotype is most widely conserved in the population. In all other respects, though, the haplotype will continue to behave in a manner identical to all other haplotypes. It will continue to be broken apart by recombination, although since selection acts upon it, its length ceases to give as reliable a reckoning of its time since origination as that of a strictly selectively neutral haplotype.

4.1.3 EXISTING METHODS OF DETECTING SELECTION

While the genetic phenomena described in the previous section have been known to exist for a considerable time, the density of polymorphism data from a wild population has rarely been sufficient to allow the practical genome-wide detection of selective sweeps by searching directly for haplotypes in which this is happening. Nonetheless, the general principles of the detection of sweeps have frequently been applied more modestly to specific loci containing genes of likely scientific interest, and have remained essentially the same for at least twelve years.

The major goal of this project was to detect instances within the UK population of *A. thaliana* of adaptation to local habitat despite gene flow from other sources, and to attempt prediction of possible cause(s). The typical pattern of a study of local adaptation involves first identifying samples displaying phenotypes of high fitness exclusively in their native habitat, and then seeking evidence that genes possessing variation associated with these traits have undergone selection in the observable past. This chapter essentially sought to reverse that process, by identifying genomic loci in samples taken from particular habitats exhibiting signatures of selection, which should serve as targets for validation via future field experiments.

Sabeti *et al.* (Sabeti et al. 2002) demonstrated an approach and thought process that served as a major inspiration for the work carried out in this chapter. Working with two loci in the human genome suspected to possess variation associated with resistance to malaria, Sabeti *et al.* (2002) identified core

haplotypes and measured the degree of conserved co-segregating similarity in flanking loci in order to estimate the age of the haplotype. Recently emerged haplotypes (those with a high degree of co-segregating variation) found at a high frequency in the studied population were marked as likely candidates for selection, having risen to high frequency before meiotic recombination broke down the linkage disequilibrium with the surrounding variation. This is unlikely for selectively neutral variation. To gauge the likelihood of any such instance being a true signature of selection, the degree of deviation from simulated haplotypes under a coalescent process was quantified. Several haplotypes were identified as exhibiting a highly significant deviation from coalescent expectations, and thus as probable instances of alleles favoured by selective sweeps.

Detection of selection across broader sections of the genome from genetic data has historically proved much more problematic, however. Genome-wide detection of selection had been attempted by comparison with predictions drawn from population genetic models (Hanfstingl et al. 1994; Hagenblad & Nordborg 2002; Nordborg et al. 2005), but prior to the advent of widely available whole genome sequencing these attempts were plagued by a lack of cross-compatibility of data from various experiments, and by difficulties in determining statistical significance due to confounding from drift and demographic factors (see Chapter **1.3.1**; for review, see (Sabeti et al. 2006)).

Methods for detecting signatures of selection fall into at least five different classes, each searching for distinct genomic patterns arising as a consequence of selective sweeps, and each with their own strengths and weaknesses. The suitability of each class of analysis to the goals of this project will now be discussed.

- Proportion of functional to non-functional mutations (Ka/Ks ratio)

Most evolution of a genotype is expected to proceed through neutral changes (*i.e.,* those with no effect on a phenotype). In terms of base substitutions within a sequence, this means that substitutions producing no change in phenotype may typically expected to be observed much more frequently than substitutions producing a change in phenotype. This ratio may be quantified by comparing the number of sequence differences producing codons coding for different amino acids (non-synonymous, or functional mutations) with that producing codons coding for the same amino acid (synonymous, or non-functional mutations). Once measured, this ratio may then be compared against either the equivalent ratio at the same loci in other species, the ratio at loci carefully chosen for their neutrality, or the typical ratio across the rest of the genome. Sustained selection over long timescales has been identified through higher proportions of non-synonymous mutations than expected by chance (McDowell 1998; Rose 2004; Ding et al. 2007). Since it is also expected that deleterious mutations are unlikely to ever rise to a high frequency in a population (due to selection acting against them), it is reasonable to conclude that such an observation constitutes a signature of a selective sweep.

This type of analysis is routinely applied to sequence data collected from closely related species, and is best suited to analysis of strong, persistent selection pressures at a single gene's locus over many millions of years. SNP data is not ideally suited to this type of analysis, but resequencing data is, such as that from the 1001 Arabidopsis Genome project. Therefore, this method was not utilised for the primary detection of sweeps, but may be useful for secondary analysis of candidate loci.

- Local reduction of genetic diversity

As a selective sweep progresses and linked alleles are drawn towards fixation by genetic hitchhiking, the genetic diversity (*i.e.,* the number of alleles in the population) at those linked loci necessarily decreases from the typical level encountered across the rest of the genome. Selective sweeps may therefore be

recognised by a sudden and progressive drop in the genetic diversity of genotypes centred on a particular locus (Carlson et al. 2005; Sabeti et al. 2006).

Eventually, diversity at the linked loci rises again. If the sweep occurred across the entire native range of the species, diversity will rise slowly as new mutations begin to appear; if the sweep occurred only across a fraction of the species' range, though diversity at these loci may be restored more quickly as migrants reintroduce variation if, for example, it were restricted to a relatively isolated sub-population.

While classic selective sweeps decrease allelic diversity at linked loci, balancing selection has been shown to actually increase diversity (Charlesworth 2006). This provides a means of not only identifying selection, but of predicting its nature.

SNP datasets are well suited to this type of analysis, which may inform us of the nature of selection e occurring up to several hundred thousand years in the past (Sabeti et al. 2006; Pritchard et al. 2010; Hernandez et al. 2011). A simple implementation of this method was carried out in this project, and the results contrasted with other methods employed in this chapter.

- Presence of high-frequency derived alleles

Derived alleles (*i.e.,* those created by mutation of ancestral alleles) usually exist at low frequency in a population. Should these alleles be linked to an allele that undergoes a selective sweep, they will be drawn towards fixation through genetic hitchhiking. Loci undergoing selective sweeps may therefore be identified by the presence of derived alleles at unusually high frequency.

This analysis requires knowledge of a population's ancestral alleles, in order that they may be distinguished from derived alleles. In *A. thaliana*, ancestral genotypes cannot be inferred with any confidence, since the population

structure and degree of admixture render any attempt futile (see Chapter 2); therefore, this method of detecting selection was not used in this project.

- Population differentiation

If a population is divided into relatively distinct sub-populations, then large differences in allele frequencies between populations may be indicative of a selective sweep (Kreitman 2000; Sabeti et al. 2007). Distinguishing the precise cause of observations of this nature in the absence of additional information is often extremely challenging, however, as the same observations may very often be attributed with at least equal plausibility to demographic effects.

Since research in this chapter set out explicitly to develop a means of distinguishing between demographic and selective effects, this method was not employed.

- Haplotype length

Loci undergoing a selective sweep are likely to maintain linkage with nearby alleles as the sweep progresses (as described in the previous section). Loci under selection are therefore identifiable due to the preservation of a greater degree of linkage than expected for their observed frequency.

Detection of selection via haplotypes may only detect very recent selection events, since large haplotypes tend to break down rapidly. On the other hand, the haplotype-based detection method is capable of detecting partial sweeps (in which the allele under selection rises in frequency, but does not reach fixation), and is relatively unaffected by any potential biases arising from choice of SNPs to use in the analysis (see Chapter **2.3.1**). This method of detection is therefore both ideally suited to the data available to this project, and to its goals.

4.1.4 DISEASE RESISTANCE IN A.THALIANA: MODEL PLANT MEETS MODEL SYSTEM

It has long been known that host-parasite interaction is a hotbed of evolutionary activity – an arms race between host and pathogen. More recently, scientists have also come to recognise host-parasite interaction as a system capable of granting insight into ecological phenomena. The main aim of this chapter was to continue this line of research by identifying loci exhibiting signatures of selection indicating that they are becoming adapted to local habitat conditions.

In order to initiate testing for the action of selection on genes affecting particular traits, as per Hypothesis **H5**, it is necessary to first construct a list of genes possessing variation known to affect the trait of interest. This may be achieved by building a list of genes reported to possess relevant qualities in published literature (as described in this section), and through wet-bench experimentation for validation of predicted loci (described in Chapter **4.3.1**).

Since the study of plant-pathogen interactions is both highly informative of evolutionary conflicts (arms races, evolutionary tradeoffs and costs of adaptation) and easily undertaken in a controlled laboratory environment, a considerable base of published knowledge has accumulated over the past several decades of research (see Chapter **1.7** for an overview and description of our current best model of plant-pathogen interactions and co-evolution). This includes many genes identified as being involved in defence against infection by specific pathogens. In some cases their methods of action are known, but this is not always the case, because knowledge is usually restricted to the general functional and perhaps structural categorisation of the gene. In addition to individual genes found to associate with model pathogens in lab experiments, then, structural gene classes (i.e., genes containing leucine-rich repeat (LRR) domains, on the grounds that the majority of *R* genes and many PRR genes contain this motif (Dangl & McDowell 2006) containing genes associated with disease resistance were also identified from the literature.

Resistance of *A. thaliana* to a highly specialised oomycete pathogen *Hyaloperonospora arabidopsidis* has been an experimental 'model system' for studying disease resistance (for review, see (Holub 2008)). For example, variation across 20 *R* genes, including the much-studied *RPP13* (Bittner-Eddy et al. 2000; Rose 2004; Charlesworth 2006), has been found to confer resistance to races of this pathogen (Nemri et al. 2010). Like many *R* genes, *RPP13* possesses an LRR region of the 'coiled coil or leucine zipper' subclass (for structural description, see (Landschulz et al. 1988). Initially, Bittner-Eddy *et al.* (Bittner-Eddy et al. 2000) interpreted allelic variation of *RPP13* as evidence of diversifying (and therefore positive) selection; whereas Rose *et al.* (Rose 2004) interpreted the variation as being maintained through balancing selection.

RPS2 is an LRR-type PRR gene that encodes a receptor protein that directly interacts with bacterial flagellin, forming part of the PAMP-triggered immunity that acts as an initial defence against infection by *Pseudomonas syringae* and other bacterial pathogens (see Chapter **1.7**). As with *RPP13*, selection favouring the long-term maintenance of allelic diversity at this locus (i.e., balancing selection) has been reported (Mauricio et al. 2003).

RPM1 is also a NB-LRR-type *R* gene in *A. thaliana*, which acts to prevent infection by the pathogenic bacterium *P. syringae*. Its R protein is known to act effectively against pathogens utilising two distinct mechanisms of host recognition through induction of rapid cell death (a hypersensitive response) (Grant et al. 1995; Boyes et al. 1998). Due to its localisation within the cytosol of the host cell and lack of transmembrane domains, it is probable that as per the 'guard hypothesis', this R protein does not interact directly with the effector produced by the pathogen but instead detects changes in host genes brought about by effector-induced inhibition. In such a case, it is reasonable to hypothesise that the *R* gene will be under either purifying or balancing selection (see Chapter **1.7.2**). Rose *et al.* (Rose et al. 2012) have recently shown that *RPM1* has been undergoing mutation and maintenance of recurrent loss-of-

function alleles in the UK and global populations, demonstrating that an ancient balancing polymorphism is not the case as first proposed (Stahl et al. 1999). The host gene guarded by *RPM1* potentially faces an evolutionary trade-off: genotypic variation leading to structural alterations in its protein may prevent the pathogen from successfully initiating an infection, but this variation is also likely to reduce the effectiveness of the R protein-mediated response. Evolution may thus either impose stabilising selection on both the guarded gene and the *R* gene, or alternatively may bring about a stable equilibrium of balancing selection in the membrane protein and a matched pattern of balancing selection in the *R* gene.

Besides the receptor-like NB-LRRs, there are many other structural classes of *R* genes, which may be involved in functions as diverse as neutralisation and degradation of pathogen toxins, kinase-mediated signaling via phosphorylation, or detection of molecular signals associated with pathogen-induced damage. Gene products carrying out these functions may be localised in the cell cytosol, in the membrane, or even in extracellular space.

While the method put forth in this chapter is suitable for detection of signatures of selection across the specific loci described above, it is also suitable for testing further hypotheses regarding selection in the context of disease resistance – namely, in testing whether specific pathogens exert selective pressures at all. This project aimed to test for signatures of selection at loci associated with resistance to the pathogen *Albugo candida* in *A. thaliana*, and in doing so, determine whether selection in the wild population is driven by this pathogen.

Albugo candida is an oomycete pathogen of a broad range of hosts across the Brassicaceae, Cleomaceae and Capparaceae (Choi et al. 2009). There is reason to suspect that infection by *A. candida* may exert a significant selection pressure on wild populations of *A. thaliana*: lab experiments show that *A. candida* is fully capable of parasitising *A. thaliana*, though *A. thaliana* also displays a substantial frequency of resistance phenotypes in such experiments. If maintenance of

resistance alleles carries a selective cost in the absence of an infective pressure, then this would imply that there is a selective force favouring any alleles conferring resistance to *A. candida* parasitism.

4.1.5 ABIOTIC CANDIDATES FOR SELECTION IN A. THALIANA

Since the analysis developed in this chapter was as capable of identifying signatures of selection arising from abiotic as well as biotic factors, the literature was also consulted regarding genes known to be associated with variation in fitness across different habitat types.

A classical example of adaptation to stressful abiotic conditions is that of heavy metal tolerance. Due to the variety of possible metal contaminants and continuous range of variation in degree of contamination, the genetics of the evolution of heavy metal tolerance has proved almost as complex as that surrounding disease resistance (Macnair 1993). While the data available to this project does not include records of degree of environmental contamination, a small number of samples within the dataset were taken from sites adjacent to railway lines – sites where, in all likelihood, industrial pollution is higher than average. Consequently, a degree of adaptation to heavy metals or other pollutants may be expected in populations at these sites.

Given the range of climatic variation between Northern and Southern regions of the UK, and given the additional range of sources of genotypic variation in the UK (see Chapter **3.3.1**) – ranging from Scandinavian to Mediterranean latitudes – there may be scope for adaptation in terms of flowering time (Alonso-Blanco & Koornneef 2000; Michaels et al. 2003).

4.1.6 OTHER KNOWLEDGE REQUIRED FOR SELECTION ANALYSIS

Several datasets either produced or utilised in earlier chapters are also incorporated in this analysis, including the haplotype dataset (see Chapter **2.3.2**). The 'hit' dataset is optional, but as with the PopAger implementation of an *in silico* population described in Chapter **3**, its use as the source of genotypes

for simulated plants is recommended. SelectionFinder relies upon reconstructing haplotypes from a population simulated under strictly selectively neutral conditions via the same method used to catalogue haplotypes in the wild population (see Chapter **2.3.4**). In order that an unbiased set of expectations may be generated for the observations drawn from the wild population, it is recommended that the same data types be used.

In common with PopAger, effective population size (N_e) and the related values for appropriate scaling of migration likelihood are also required. Likewise, meiotic crossover events per chromosome and outcrossing rate should be scaled to control the degree of drift experienced by the simulated population (see Chapter **3.3.3**).

Clustering based on Structure analysis, such as that performed in Chapter 2, may be used in place of straight distance between sample collection sites if desired. This may result in a better representation of the likelihood of outcrossing in a population following a more complex structure than that reported for *A. thaliana*, but makes distinguishing signatures of selection more difficult, since Structure clustering inherently incorporates any signatures of selection present in the genotype dataset supplied to it.

4.2 MATERIALS AND METHODS

4.2.1 MAGIC QTL MAPPING

Pathogen lines were stored long-term at -80C as asexual inoculum in frozen leaf tissue. Prior to inoculation, they were grown into a bulk stock in a broadly susceptible accession of *A. thaliana* (Wassilewskija-2)(see (Borhan et al. 2001).

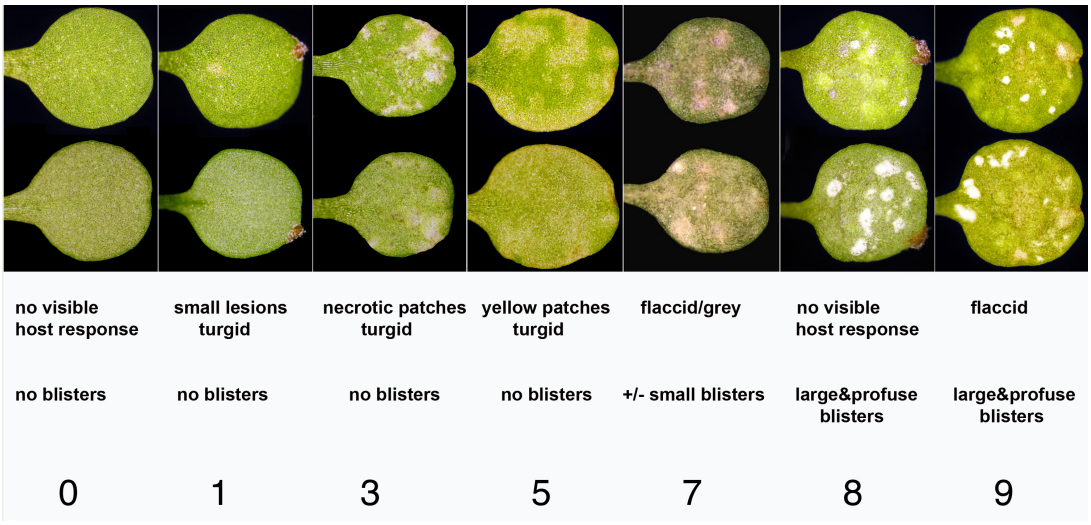


Figure 22 Interaction phenotype scale for response of *Arabidopsis thaliana* following infection with *Albugo candida*. Host responses were graded on a semi-quantitative scale, ranging from complete inability of the pathogen to cause symptoms or reproduce to full susceptibility (indicated by the profuse formation of white blisters containing zoosporangia).

QTL analysis was performed using three isolates of *A. candida* on 405 of the *A. thaliana* MAGIC lines developed by Kover *et al.* (Kover et al. 2009). Seeds from the 19 parent lines and each of the RILs were sown ca. 10 per pot in a Levingtons F2 soil selected for optimal growth of *A. thaliana,* and cold treated at 8^{0}C for five days in the dark. The trays of sown seed were then transferred to a controlled environment room (16h day at 150 microEinsteins, 8h night). The RILs were inoculated after 10-12 days of growth by suspending *A. candida* zoosporangia in distilled water to a concentration of $5x10^{4}$ zoosporangia per ml, and then spraying the suspension over the pots, following appropriate sterile procedure. The inoculated plants were then incubated in the same growth room for a further 10-12 days before their interaction phenotypes were observed and recorded. Phenotypes were scored using a semi-quantitative

scale (shown in Figure **22**), enabling investigation of phenotypes showing an incursion of the pathogen into the host leaf tissue, but a host response sufficient to prevent a full infection leading to sporulation. These phenotypes were later converted to binary classes (Resistant = IP classes 0-7; Susceptible = IP 8-9). Observation of phenotypes was possible with the naked eye, though frequently easier with the aid of a magnifying lens.

Loci were mapped to differences in phenotype by identifying parts of the genome showing co-segregation with the observed phenotypes. This analysis was carried out in R, using the HAPI package developed by Kover *et al*.. (2009)

4.2.2 SELECTIONFINDER

Much of the programming required to produce the SelectionFinder tool itself and the supplemental analysis tools developed alongside it was carried out, as with other work in this project, in Perl. Many components of the SelectionFinder tool use code similar or identical to the PopAger tool developed in Chapter 3 – particularly, those involved in simulating the distribution and propagation of individual seeds into reproducing plants.

Geographic distances between sample collection sites were calculated using the Haversine method using module GIS::Distance. Analysis of genetic similarity between samples using the program Structure may also be used in place of this data. Any situation requiring differences or similarities between two lists was handled by module List::Compare. Various common mathematical functions were carried out by module Math::NumberCruncher. Number of crossovers in any given simulated outcrossing event was given a realistic variation in range (a normal distribution centred around the supplied mean value of three per chromosome) using module Math::Random::OO:Normal. The descendent genotypes were assembled from the chosen pattern of crossover points using module IntervalTree. Weighted random choices, including the sites to disperse newly-created seeds to, and the loci at which crossovers were to occur in an outcrossing event, were carried out using module List::Util::WeightedChoice. As

with the original haplotype generation process in Chapter 2, haplotypes were recreated at the end of the analysis with a clustering-based approach utilizing modules Algorithm::Cluster and Algorithm::Cluster::Thresh to carry out hierarchical clustering and thresholding of that clustering respectively.

Data showing the precise locations and functional categories of genes within the *A. thaliana* genome was taken from TAIR. Data detailing the variation in crossover rates across the genome was kindly given by M. Horton.

A comparison between genomic regions identified by SelectionFinder as exhibiting a signature of selection and SNP diversity within the sub-population in question was also implemented in perl using basic mathematical functions from module Math::NumberCruncher, and plotted using module ImageMagick. Functional data relating to genes within these genomic regions was taken from the 1001 Genomes Project (Weigel & Mott 2009). GO terms assigned to these genes were also used (Gene Ontology Consortium 2004).

Ka/Ks ratios were measured as a simple ratio by counting synonymous vs. non-synonymous SNPs in 1001 Genomes data gathered from UK accessions used in the 250K dataset.

All processing for this analysis was carried out on a 2009-built dual-core MacBook Pro with 8GB RAM.

4.3 RESULTS

4.3.1 QTL ANALYSIS OF A. CANDIDA RESISTANCE IN A. THALIANA

A substantial range of phenotypic variation was observed in response to challenges from *A. candida*, from complete susceptibility, through several degrees of incomplete susceptibility in which the parasite is able to enter the leaf tissues but unable to sporulate, to complete resistance (see Figure **22**).

Table 5 QTLs from analysis of pseudo-quantitative phenotype observations

Chromosome	From (bp)	To (bp)	peak.bp	peak.SNP	logP	Genome-wide p-value
1	NA	26405125	20803191	MASC00513	51.16437677	0

Table 6 QTLs from analysis of binary (resistant/susceptible) phenotype observations

Chromosome	From (bp)	To (bp)	peak.bp	peak.SNP	logP	Genome-wide p-value
1	18228436	21941097	20803191	MASC00513	7.466611783	0
1	22279846	22350442	22286231	MASC04170	3.531431563	0.04
5	4428858	6523118	5829586	TFL2_775	5.176057663	0.002

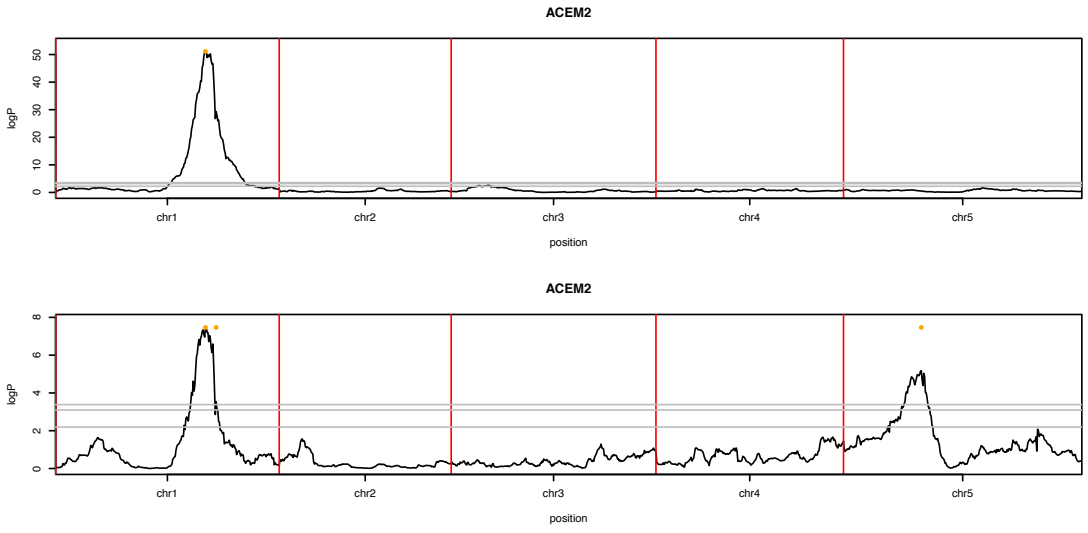


Figure 23 ***A. candida*** **Infection QTL traces** from analysis of pseudo-quantitative phenotypic observations (**ABOVE**), and binary resistant/susceptible observations (**BELOW**). Co-segregation of WRR4 (Chromosome 1) is strongly significant across both sets of observations, but segregation of the WRR5/6 locus (Chromosome 5) becomes apparent only in binaryised (semi-quantitative scale data converted to susceptible/resistant format) data.

Phenotypes recorded from MAGIC lines exposed to infection by the ACEM2 isolate of *A. candida* (Ploch et al. 2010) were analysed using software developed in conjuction with the MAGIC mapping method (Kover et al. 2009). Phenotypes were submitted for analysis in both a binary (susceptible/resistant) and pseudo-quantitative format. The analysis reported QTLs corresponding to the *WRR4* locus (Borhan et al. 2001; Borhan et al. 2008; Holub 2008) on chromosome 1 and the *WRR5/WRR6* gene pair on chromosome 5 (Cevik & Holub, unpublished). QTLs are shown in Tables **5** and **6**, and Figure **23**.

4.3.2 SELECTIONFINDER TOOL

In order to check for haplotypes currently undergoing selective sweeps, a tool provisionally named SelectionFinder was created. SelectionFinder works on similar principles to a GWAS, in that it implements a statistical comparison of allele frequencies between two groups of samples. It differs from a standard GWAS, however, in that its 'control' population is derived from a model population under conditions of total selective neutrality, and also in its ability to identify alleles differing from typical distribution trends in spatial dimensions. There is also a facility to divide the 'case' group – the real population – into smaller groups (e.g., separating the regional population into smaller groups based on habitat), which may independently be compared against neutral expectations. The aim of this analysis is to detect haplotypes in the wild population exhibiting a statistically significant difference from either demographic or spatial trends observed in the neutral model population, thus separating linkage disequilibrium attributable to selection from that attributable to other factors representing deviations from Hardy-Weinberg equilibrium.

SelectionFinder generates its range of neutral expectations using a population simulation approach similar to that developed in Chapter **3**. A simulation-based approach was again chosen for this application since it allows for the independent control of both demographic factors and, crucially for this analysis,

selection upon genotypes. The genotypes of this *in silico* population may be supplied either as haplotypes, or as clusters identified from each of the 'windows' along the length of the genome (*i.e.,* 'hits'; see Chapter **2.3.2** for details). As in Chapter **3**, 'hit' data should typically be used, although should an analysis require the simulation of haplotypes over a number of generations, the option is present.

Unless otherwise specified, SelectionFinder analysis makes two implicit simplifying assumptions regarding the simulated population:

- That the 'isolation by distance' population structure is determined entirely by distance between sites (unless Structure data is supplied in order to attempt to mitigate this oversimplification);
- That in a neutral population, crossovers occur with equal frequency at all parts of the genome (unless crossover rate data is used to mitigate this oversimplification);

Unlike the PopAger tool in Chapter **3**, SelectionFinder makes no distinction between 'Population' and 'Background' samples, since migration from other areas was not included as a factor in this investigation. Instead, this tool simply analyses genotypes from a single area (known here as the 'active' area) – in the case of this project, the UK. Since migration from outside sources is not modeled in this analysis, genotypes recorded from samples outside the 'active' area, if any, need not be retained beyond this point. In terms of input, this means that a simulation may be very easily programmed on the sites spanning an arbitrarily selected area.

Initial measurements consist simply of reading in the haplotype dataset (and, optionally, the 'hit' dataset) produced in Chapter **2**. The first generation of plants is set up, as with PopAger, by creating a group of simulated individuals possessing genotypes drawn from either the haplotype or 'hit' datasets; unlike PopAger, however, this group of individuals represents the present-day

population within the 'active' area, so comprises all genotypes collected from within the area, rather than a small number of hypothetical founders.

The population is then allowed to propagate in the same way as that described for PopAger. Each simulated plant creates multiple seeds, either by self-fertilisation or by meiotic recombination. Selection of pairs of plants for outcrossing (representing dispersal of pollen) proceeds as previously stated, in that the second parent is selected with a weighted random choice in which the weights are determined by either geographic distance, or by clustering results from Structure analysis. Simulation parameters controlling frequency of outcrossing and the number of crossovers per chromosome are scaled to account for the discrepancy in size between the simulated and real populations in precisely the manner described in Chapter **3.3.3**.

It is again recommended that each plant is set to produce multiple seeds, to ensure the population remains at or near its maximum allowed number of individuals (and is capable of recovery if, for some reason, its numbers fall). Differential rates of crossover incidence at various points across the genome may be modeled during outcross events, though this facility was not utilised in this case (in favour of a purely random choice of crossover sites) as it was feared a differential choice of crossover sites would unduly bias some parts of the genome towards false positive and false negative readings of signatures of selection.

Seeds generated up to this point are then dispersed. The user is able to choose between basing the likelihood of dispersal from one site to another on either straight-line distance, or on measurements of genotypic similarity derived from Structure analysis. Since the latter may incorporate measurements of genotypic similarity attributable to selection rather than gene flow, it is recommended that straight-line distances be used as the basis of dispersal likelihoods. Additionally, if differential crossover rates across the genome are used during the production of seeds with recombinant genotypes, it is recommended that

the Structure analysis incorporate these data too. In this project, straight-line distance data was also generally preferred due to the wide geographic range of *A. thaliana* and large number of potential habitat sites. Both seed and pollen dispersal rates are calculated to account for the smaller size and more limited dispersal options of the simulated population, as shown in Chapter **3.2.2**. Unlike the application of the PopAger tool in Chapter **3**, however, only one set of these parameters was utilised in this chapter, as the use of other parameter sets was not considered to be sensible in this context.

Following dispersal, stationary seeds derived from the extant parent generation may also be added to the seed pool. As with PopAger, this option is presented in order that the simulated population conforms more closely to the nature of the wild *A. thaliana* population.

Seeds from the seed pool are then chosen to form the next generation of plants. Again, the user specifies the maximum number of plants that may be accepted to do so at any one site. Higher numbers of seeds per site are preferable, since this brings the influence of drift on the frequencies of alleles and the value of θ for haplotypes more closely in line with that of the wild population. However, use of dense genotype data – particularly hit data – with a large number of seeds allowed per site is likely to result in a substantial consumption of computational resources and time. Allowing more than one seed per habitable site when using genotypes assembled from 'hit' data resulted in the simulation program using excessive memory (i.e., an 'out of memory' error). For the purposes of this analysis, which requires selective neutrality in the simulated population, seeds were chosen purely randomly from the pool of seeds created at each site, until the preset number of seeds per site was reached. However, should future research require a specific selection pressure to be modelled, this may be achieved through the application of a weighted choice to the seeds present at each site.

Unlike PopAger, this generational cycle proceeds for a pre-set number of generations, before halting for a final analysis. This number of generational cycles should be at least high enough for plants at each sampling site to have begun to disperse and admix in a manner reflecting the wild population. This figure depends largely upon the number of plants allowed to develop, but also upon the outcrossing rate and number of crossovers specified per chromosome. On the other hand, despite compensations, this simulation will invariably be subject to a greater degree of genetic drift than a wild population of a large number of individuals, which causes alleles to begin to go extinct at a greater rate than occurs in the wild population as the simulation progresses. A value of 100 generations per simulation was therefore set for this analysis. Again, multiple repeats are strongly recommended, since this increases the number of data points available to each haplotype length class, and ensures better estimates of the likelihood of rare combinations of length and frequency.

Upon the completion of the final generational cycle, the genotypes of the extant simulated plants are passed to a function that reconstructs haplotypes from the simulated genotype data (see Chapter **2.3.4**). This creates a second set of haplotypes – a set drawn from a population that has propagated through several hundred generations of either selective neutrality or strictly controlled selection. From this point, the analysis proceeds along lines more similar to a more standard GWAS; the goal is to identify alleles showing statistically significant differences in frequency between the two populations.

The frequencies of alleles within the wild population cannot be directly compared against a counterpart in the simulated population in order to generate an odds ratio (as in a conventional GWAS). While corresponding haplotypes may indeed be tracked from the first generation of the simulation to the last, analysis of their precise state in the final generation of the simulation provides only limited information, since the emergence and rearrangement of new haplotypes that necessitates the re-evaluation of haplotype structure at the

end of the simulation. Instead, the length and frequency characteristics of haplotypes observed in the wild population must be compared against more general expectations derived from typical observations of those characteristics from the neutral population. Figure **24** shows the overall process of the SelectionFinder analysis.

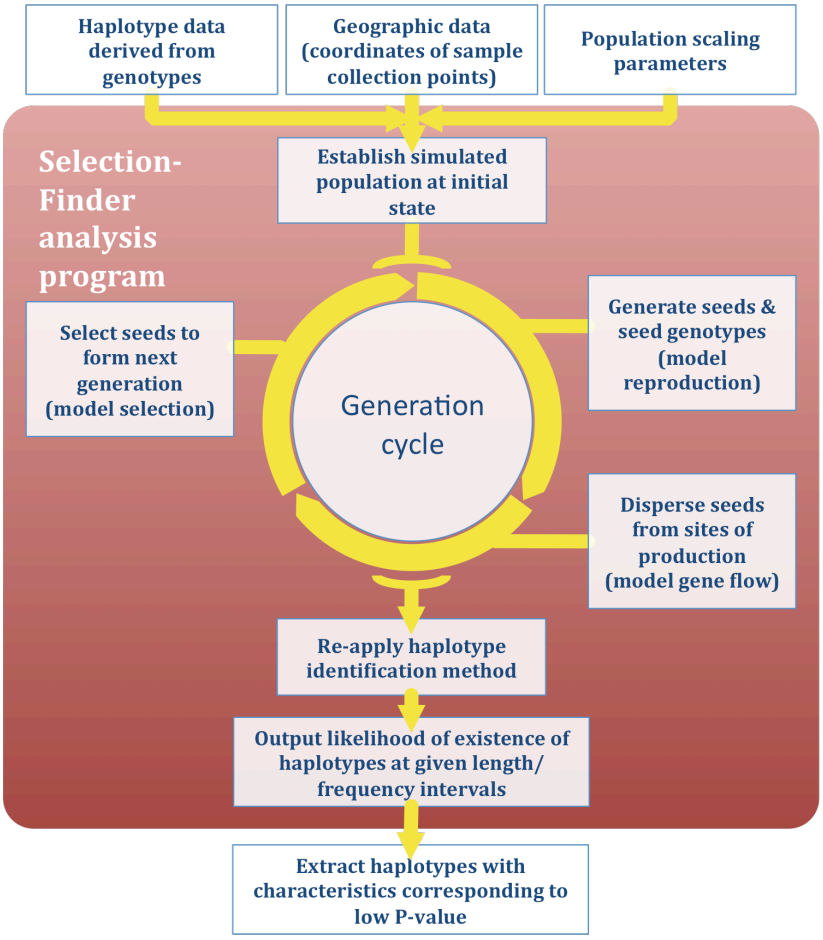


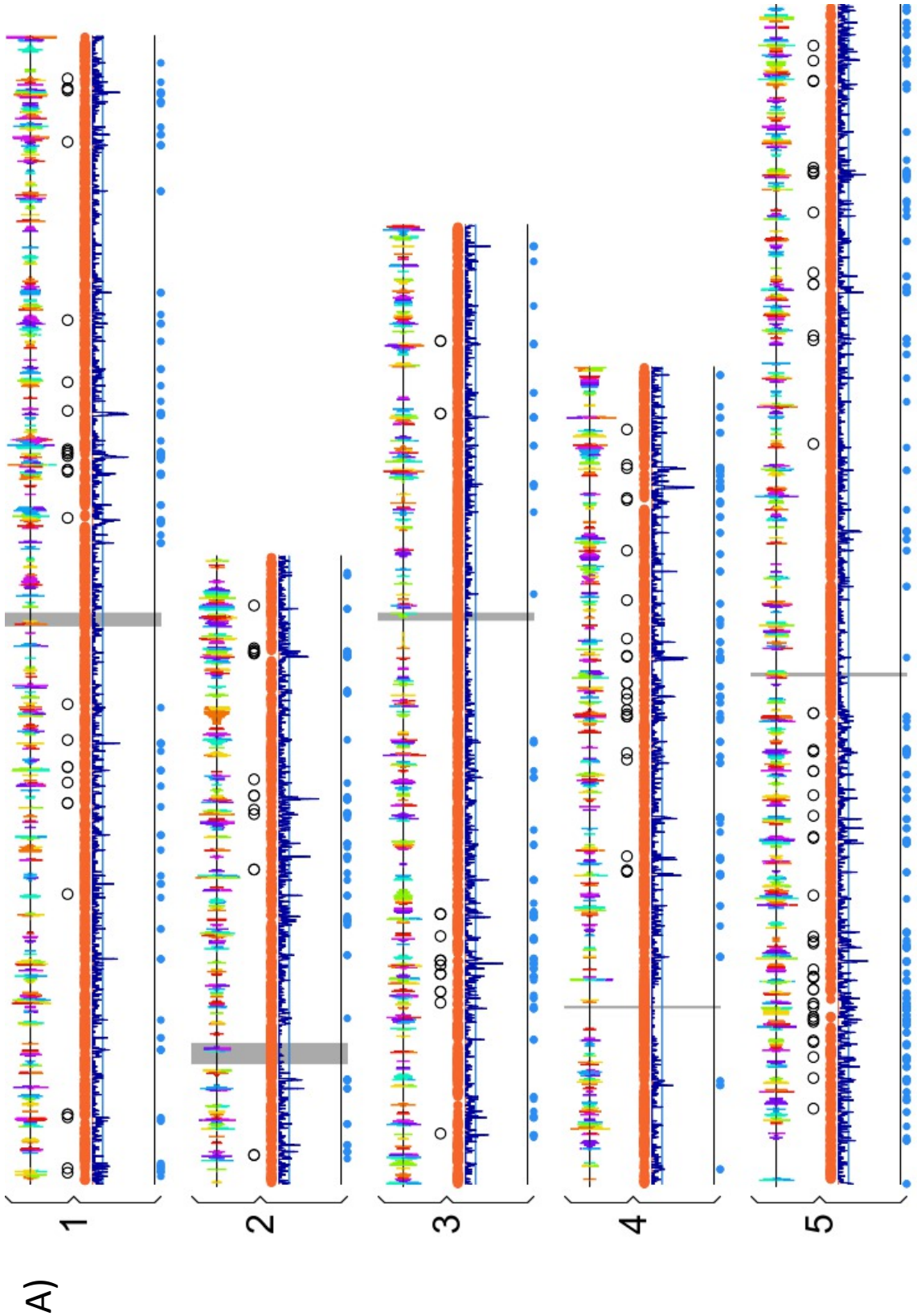
Figure 24 SelectionFinder analysis overview. Flow chart showing the SelectionFinder analysis process. This analysis aimed to generate a set of expectations based on the model of gene flow and individual dispersal developed in Chapter **3.3.2**. A simulated population was caused to propagate for 100 generations, following only the rules of gene flow set out in Chapter **3.3.2**. As with Hardy-Weinberg equilibrium, departures from this simplified model indicate the presence of additional ecological factors acting upon the population. To implement a comparison between the observed genotypes and the predictions of the model, a second set of haplotypes was collected from the genotypes of the simulated population upon completion of the simulation. The lengths and frequencies of haplotypes in this set of simulated genotypes was used to quantify the likelihood of given combinations of haplotype length and frequency (see Chapter **4.1.2** for discussion of the meaning of haplotype length and frequency) given the population dynamics of the model. Since the model accounts for much of the observed distributions and frequencies of genotypes (see Chapter **3.3.6** and Figures **12** and **21**), any haplotype possessing characteristics of frequency or genomic length significantly different from these expectations is likely to indicate the action of an ecological factor not represented in the model. Since natural selection is explicitly not represented in the model, haplotypes undergoing selection are therefore likely to possess size and frequency characteristics which distinguish them from the background of selectively neutral variation represented in the simulation.

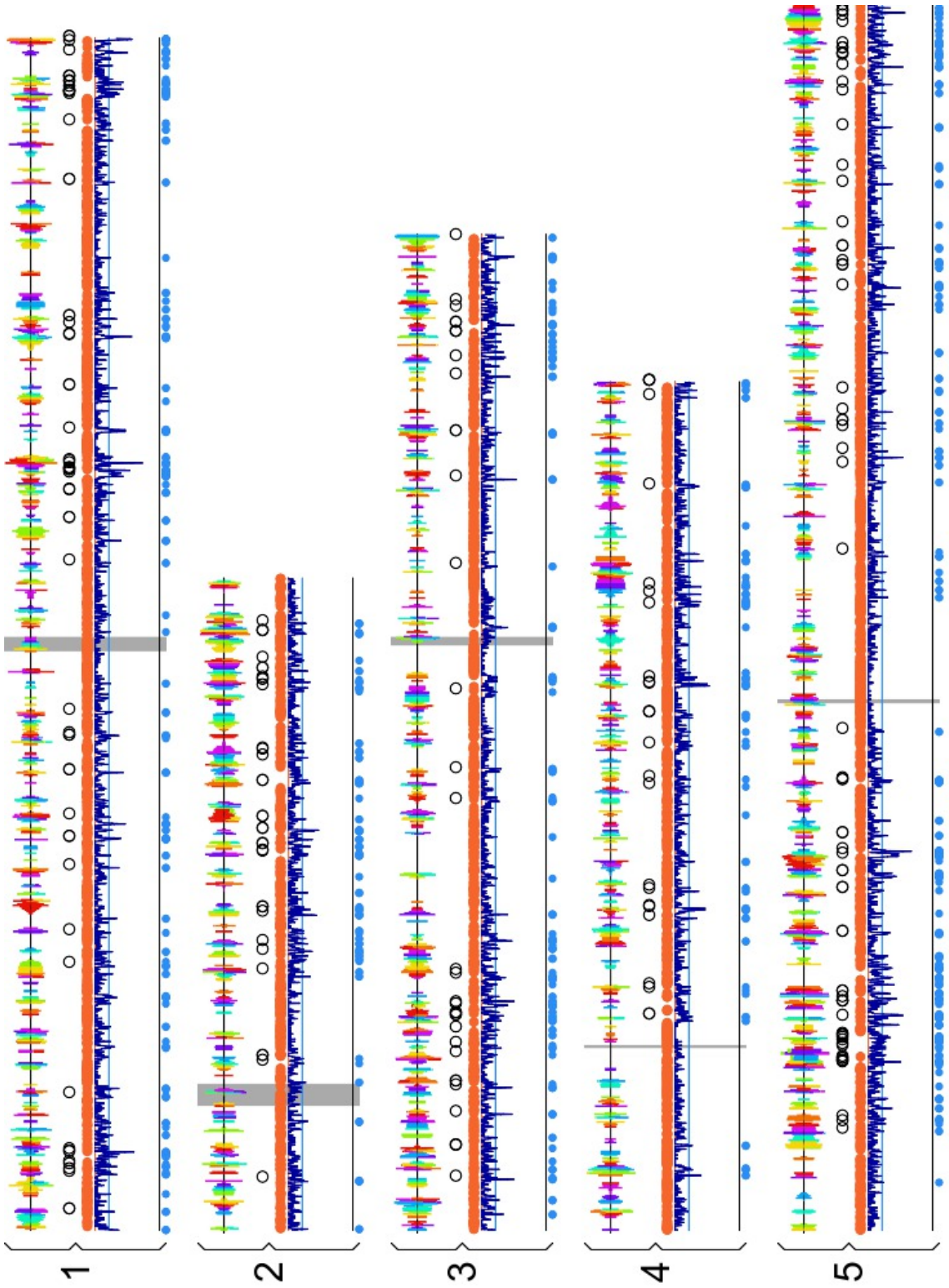
This is achieved by using the nonparametric distributions across length, frequency and spatial distribution classes of haplotypes from the neutral population as estimates of the likelihood of given combinations of these characteristics under neutrality. Most prominently, p-values for frequency

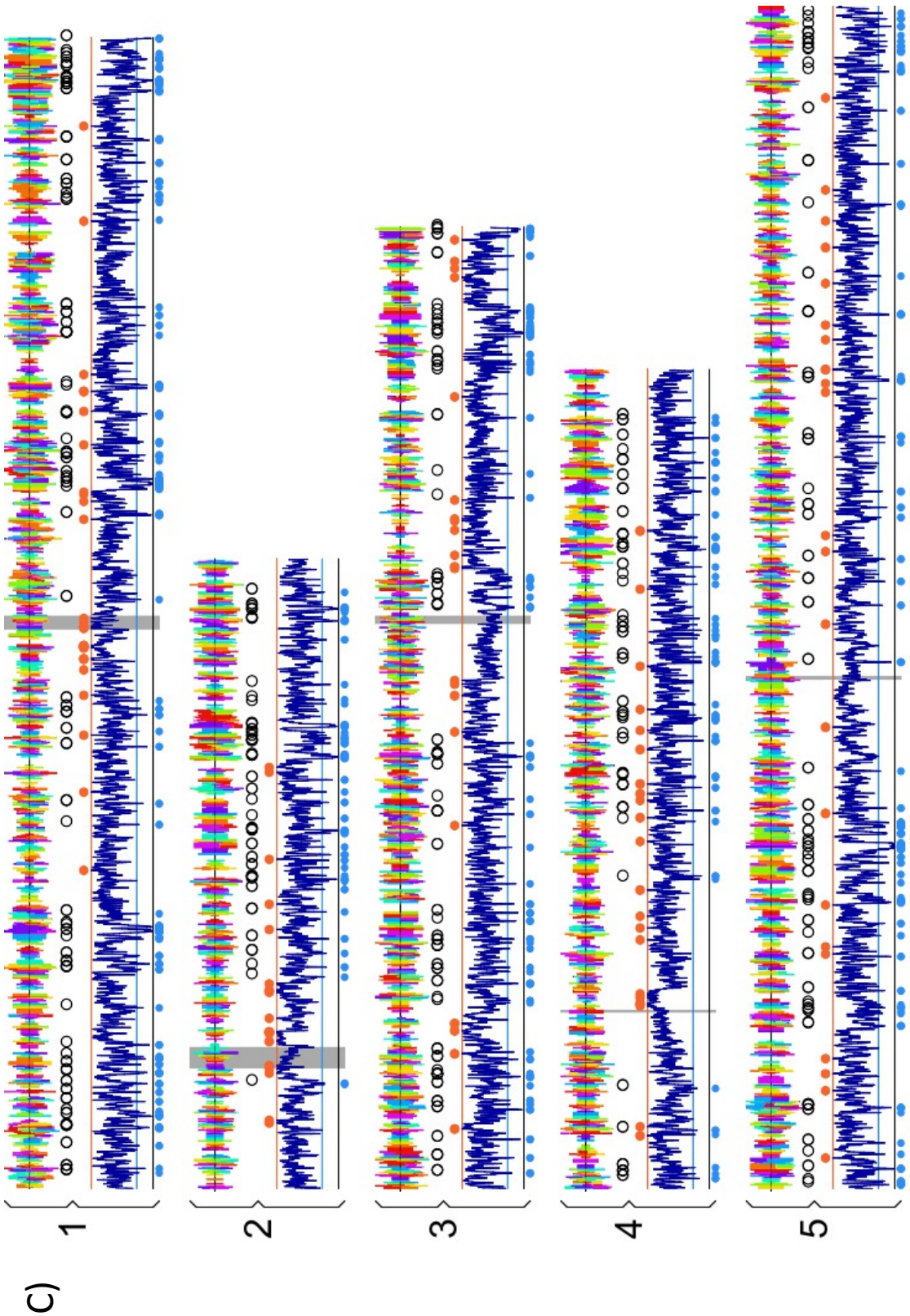
classes under neutrality were calculated for each haplotype length class. Haplotypes returning p-values beneath the threshold of statistical significance ($p<0.05$), but with frequencies beneath the mean frequencies in that length class, were excluded from further consideration on the grounds that this was unlikely to itself constitute evidence of a selective sweep.

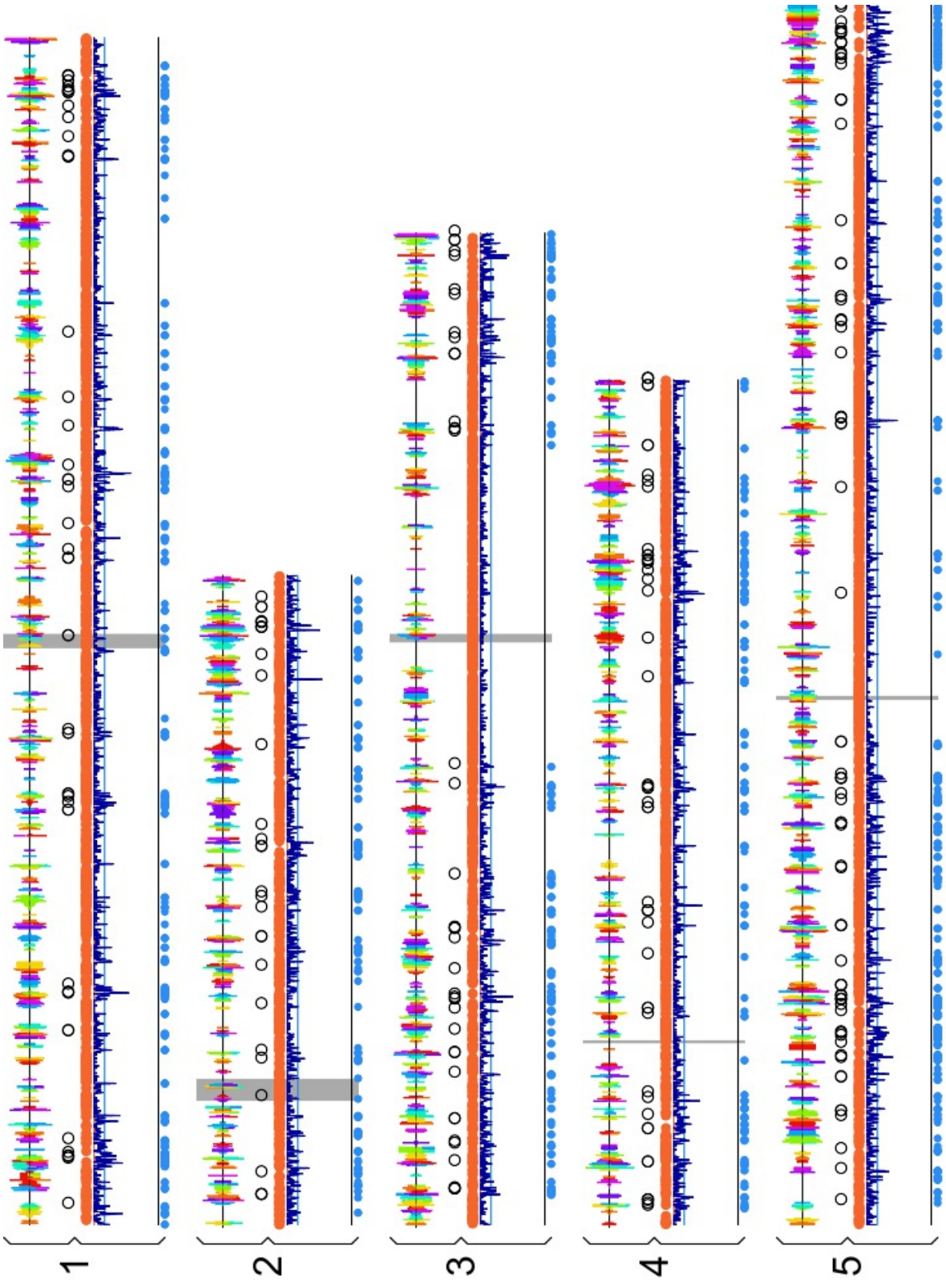
Upon the identification of haplotypes possessing demographic qualities characteristic of selective sweeps, the 'centre' of each haplotype was identified. The extent of the genome of each sample possessing the haplotype was retrieved from memory, and the centre-point of that contiguous section of genome was recorded (see Appendix **3** and supplementary data for a full list of samples possessing each marked haplotype). A haplotype's centre-point was then calculated by taking the mean of all centre-points from all individuals sharing the same haplotype. Genes at, and immediately surrounding, this centre-point locus were then retrieved from NCBI annotation data cataloguing features of the *A. thaliana* genome, along with corresponding GO terms. Finally, the SNP diversity at loci surrounding each potential signature of selection was tested, in order to identify the type of selection in each case. Loci exhibiting signatures of selection within the sample sets from each habitat type were plotted against the degree of SNP diversity at those loci (see Figure **25**).

To demonstrate the utility of this analysis, the resulting gene list (as AGI-numbers) was used for a bibliographic search to identify shortlists that may share a similar functional role in selection of a population. For example the gene list was compared against a list of candidate pathogen receptor-like genes (*i.e.,* NBS-LRR, receptor-like kinases (RLKs) and receptor-like proteins (RLPs), including specific examples described above in the Chapter Introduction), to derive a shortlist of genes that may be affected by selection pressure from


A)
1
2
3
4
5


B)
1
2
3
4
5


C)
1
2
3
4
5


D)
1
2
3
4
5

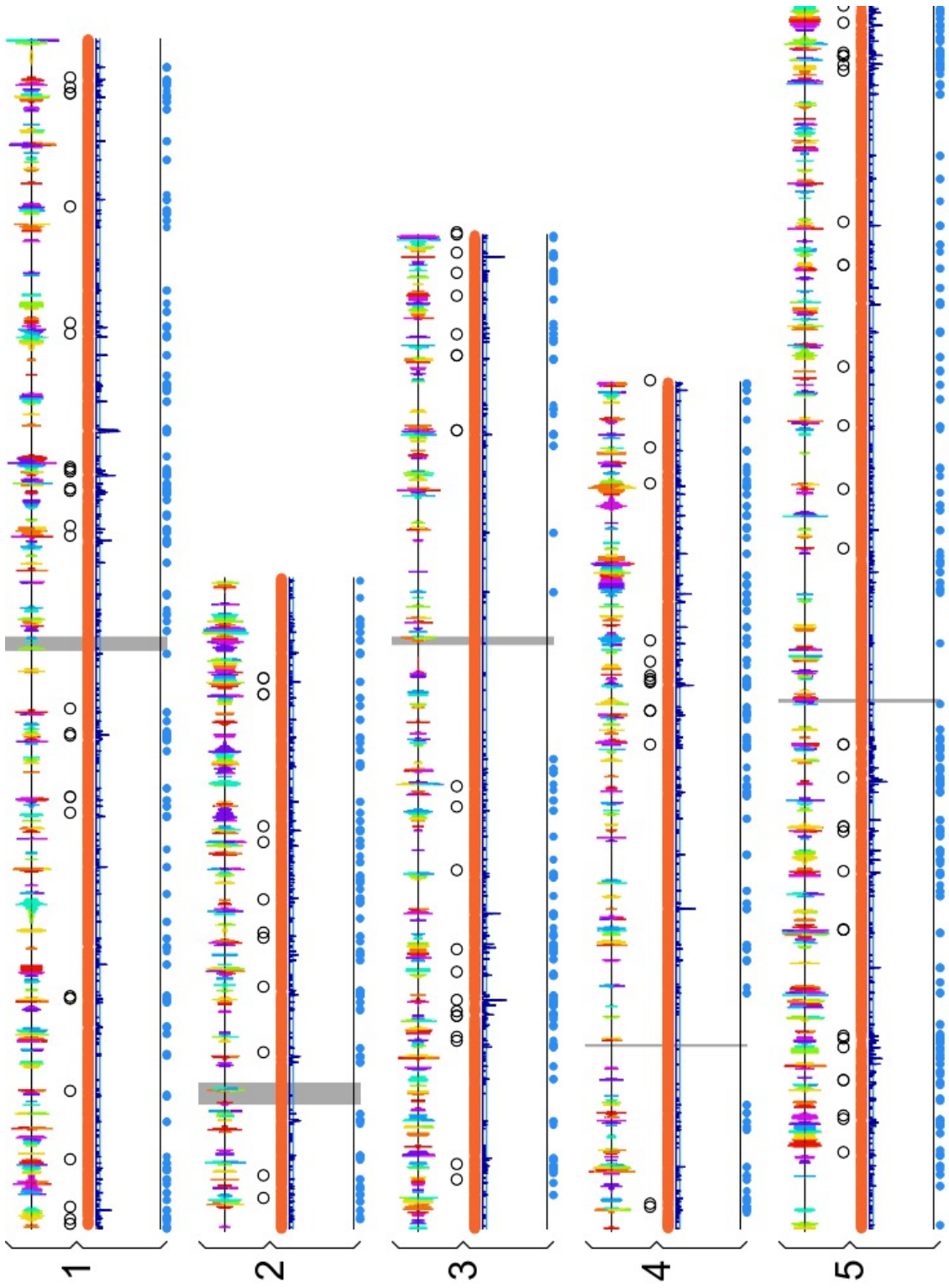

E)
1
2
3
4
5

Figure 25 (Previous pages) Haplotypes identified as departing significantly from the Neutral Model by SelectionFinder Haplotypes at an unexpected frequency in the population for their length were identified in different habitats including: **A)** wall/rocky outcrop, **B)** garden, **C)** railway and **D)** other. The same analysis was also carried out across the UK as a whole (**e**). Haplotypes across all five chromosomes are shown. The centromeric regions of each chromosome are marked in **GREY**. The plot for each chromosome is divided into two parts: an upper and lower bracket. The upper bracket indicates the locus and frequency of each haplotype identified by SelectionFinder as indicative of selection (colour is used only to distinguish adjacent haplotypes from each other). Circular markers between the brackets indicate loci that possess signatures of selection, as identified by SelectionFinder and by atypical SNP diversity. The lower bracket indicates the degree of allelic diversity across sliding windows of 30 SNPs at corresponding loci. Upper and lower significance thresholds in this series are marked; **BLUE** spots at the bottom of the allelic diversity plot mark unusually low diversity, indicating loci potentially undergoing selective sweeps. **ORANGE** spots at the top of the allelic diversity plot were intended to identify unusually high diversity, indicating loci potentially undergoing balancing selection. However, typical SNP diversity measured from available data across all habitat types except railways (**C**) was too high to allow loci possessing unusually high SNP diversity to be distinguished.

plant-microbe interactions (Appendix **6.1**), or selection pressure acting upon the phenotype of flowering time (Appendix **6.2**).

4.3.3 SELECTIONFINDER RESULTS

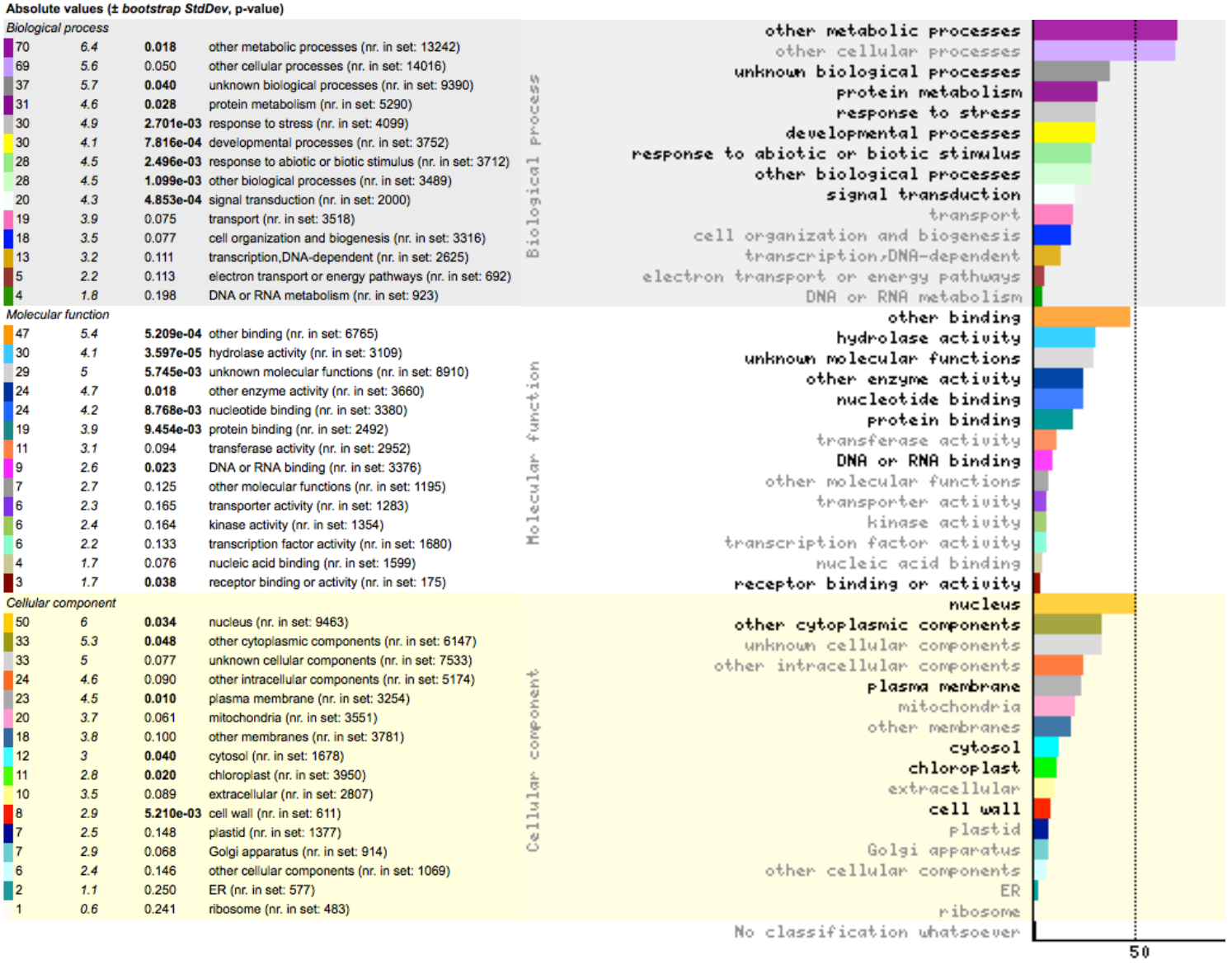
SelectionFinder analysis was applied to the UK population, and to the populations resident across four different habitat types, using the settings found to produce a good fit to the observed wild population in Chapter **3.3.5** (parameter set 2).

Histograms of the haplotype length distributions from both the real and simulated populations are shown in Figure **15**. The simulated population appears to have longer haplotypes than the actual sample from the wild population. This may be explained by the inherently greater degree of drift in the simulated population, which must be considerably smaller than the wild population, and will therefore inevitably lose rare alleles as generations progress despite measures taken to reduce the apparent effects of drift.

Also, Figure **26** shows the frequencies of cellular function classes of genes that are closest to the centres of haplotypes for each population, measured using the

classification superviewer tool from the BioAnalytic Resource toolkit (Provart & Zhu 2003). This combined analysis provides a summary of candidate genes

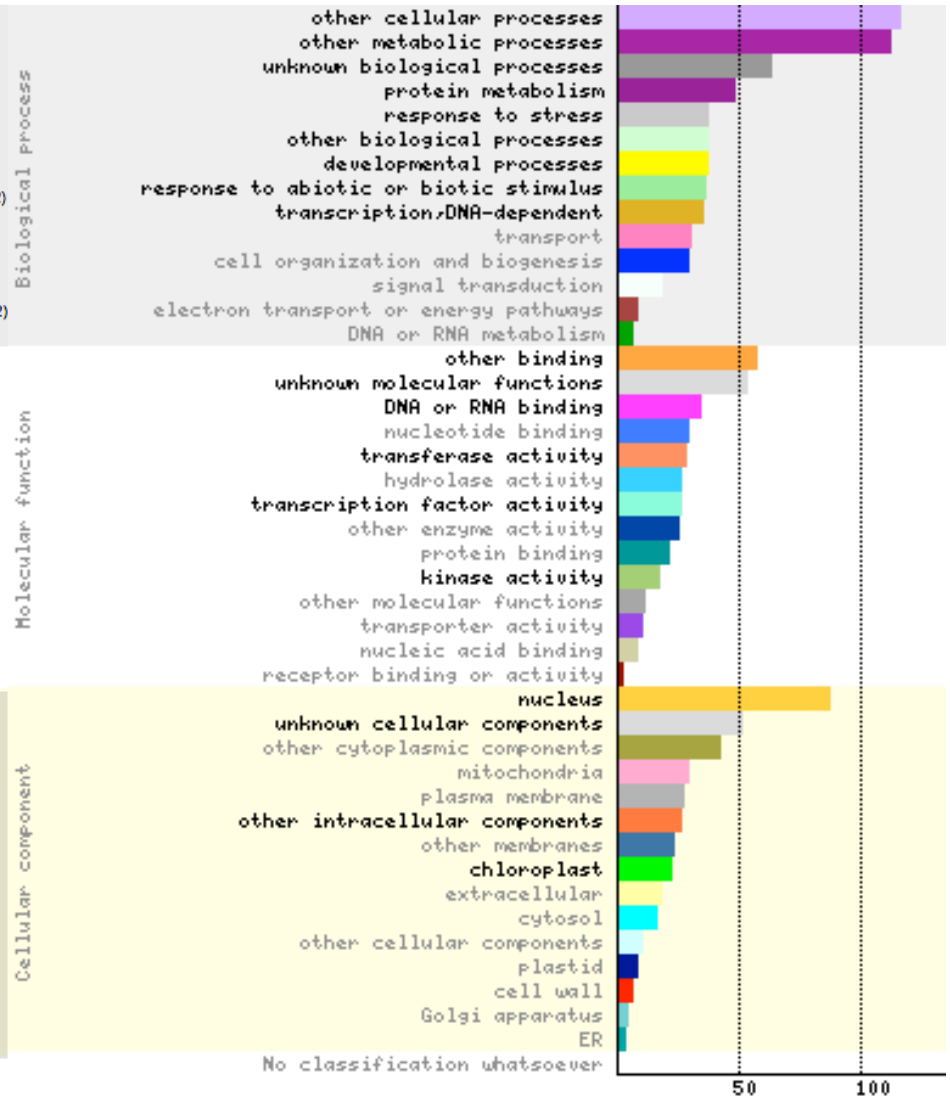
A)


Absolute values (± bootstrap StdDev, p-value)
Biological process
70 6.4 0.018 other metabolic processes (nr. in set: 13242)
69 5.6 0.050 other cellular processes (nr. in set: 14016)
37 5.7 0.040 unknown biological processes (nr. in set: 9390)
31 4.6 0.028 protein metabolism (nr. in set: 5290)
30 4.9 2.701e-03 response to stress (nr. in set: 4099)
30 4.1 7.816e-04 developmental processes (nr. in set: 3752)
28 4.5 2.496e-03 response to abiotic or biotic stimulus (nr. in set: 3712)
28 4.5 1.099e-03 other biological processes (nr. in set: 3489)
20 4.3 4.853e-04 signal transduction (nr. in set: 2000)
19 3.9 0.075 transport (nr. in set: 3518)
18 3.5 0.077 cell organization and biogenesis (nr. in set: 3316)
13 3.2 0.111 transcription,DNA-dependent (nr. in set: 2625)
5 2.2 0.113 electron transport or energy pathways (nr. in set: 692)
4 1.8 0.198 DNA or RNA metabolism (nr. in set: 923)
Molecular function
47 5.4 5.209e-04 other binding (nr. in set: 6765)
30 4.1 3.597e-05 hydrolase activity (nr. in set: 3109)
29 5 5.745e-03 unknown molecular functions (nr. in set: 8910)
24 4.7 0.018 other enzyme activity (nr. in set: 3660)
24 4.2 8.768e-03 nucleotide binding (nr. in set: 3380)
19 3.9 9.454e-03 protein binding (nr. in set: 2492)
11 3.1 0.094 transferase activity (nr. in set: 2952)
9 2.6 0.023 DNA or RNA binding (nr. in set: 3376)
7 2.7 0.125 other molecular functions (nr. in set: 1195)
6 2.3 0.165 transporter activity (nr. in set: 1283)
6 2.4 0.164 kinase activity (nr. in set: 1354)
6 2.2 0.133 transcription factor activity (nr. in set: 1680)
4 1.7 0.076 nucleic acid binding (nr. in set: 1599)
3 1.7 0.038 receptor binding or activity (nr. in set: 175)
Cellular component
50 6 0.034 nucleus (nr. in set: 9463)
33 5.3 0.048 other cytoplasmic components (nr. in set: 6147)
33 5 0.077 unknown cellular components (nr. in set: 7533)
24 4.6 0.090 other intracellular components (nr. in set: 5174)
23 4.5 0.010 plasma membrane (nr. in set: 3254)
20 3.7 0.061 mitochondria (nr. in set: 3551)
18 3.8 0.100 other membranes (nr. in set: 3781)
12 3 0.040 cytosol (nr. in set: 1678)
11 2.8 0.020 chloroplast (nr. in set: 3950)
10 3.5 0.089 extracellular (nr. in set: 2807)
8 2.9 5.210e-03 cell wall (nr. in set: 611)
7 2.5 0.148 plastid (nr. in set: 1377)
7 2.9 0.068 Golgi apparatus (nr. in set: 914)
6 2.4 0.146 other cellular components (nr. in set: 1069)
2 1.1 0.250 ER (nr. in set: 577)
1 0.6 0.241 ribosome (nr. in set: 483)
Biological process
other metabolic processes
other cellular processes
unknown biological processes
protein metabolism
response to stress
developmental processes
response to abiotic or biotic stimulus
other biological processes
signal transduction
transport
cell organization and biogenesis
transcription,DNA-dependent
electron transport or energy pathways
DNA or RNA metabolism
Molecular function
other binding
hydrolase activity
unknown molecular functions
other enzyme activity
nucleotide binding
protein binding
transferase activity
DNA or RNA binding
other molecular functions
transporter activity
kinase activity
transcription factor activity
nucleic acid binding
receptor binding or activity
Cellular component
nucleus
other cytoplasmic components
unknown cellular components
other intracellular components
plasma membrane
mitochondria
other membranes
cytosol
chloroplast
extracellular
cell wall
plastid
Golgi apparatus
other cellular components
ER
ribosome
No classification whatsoever
50

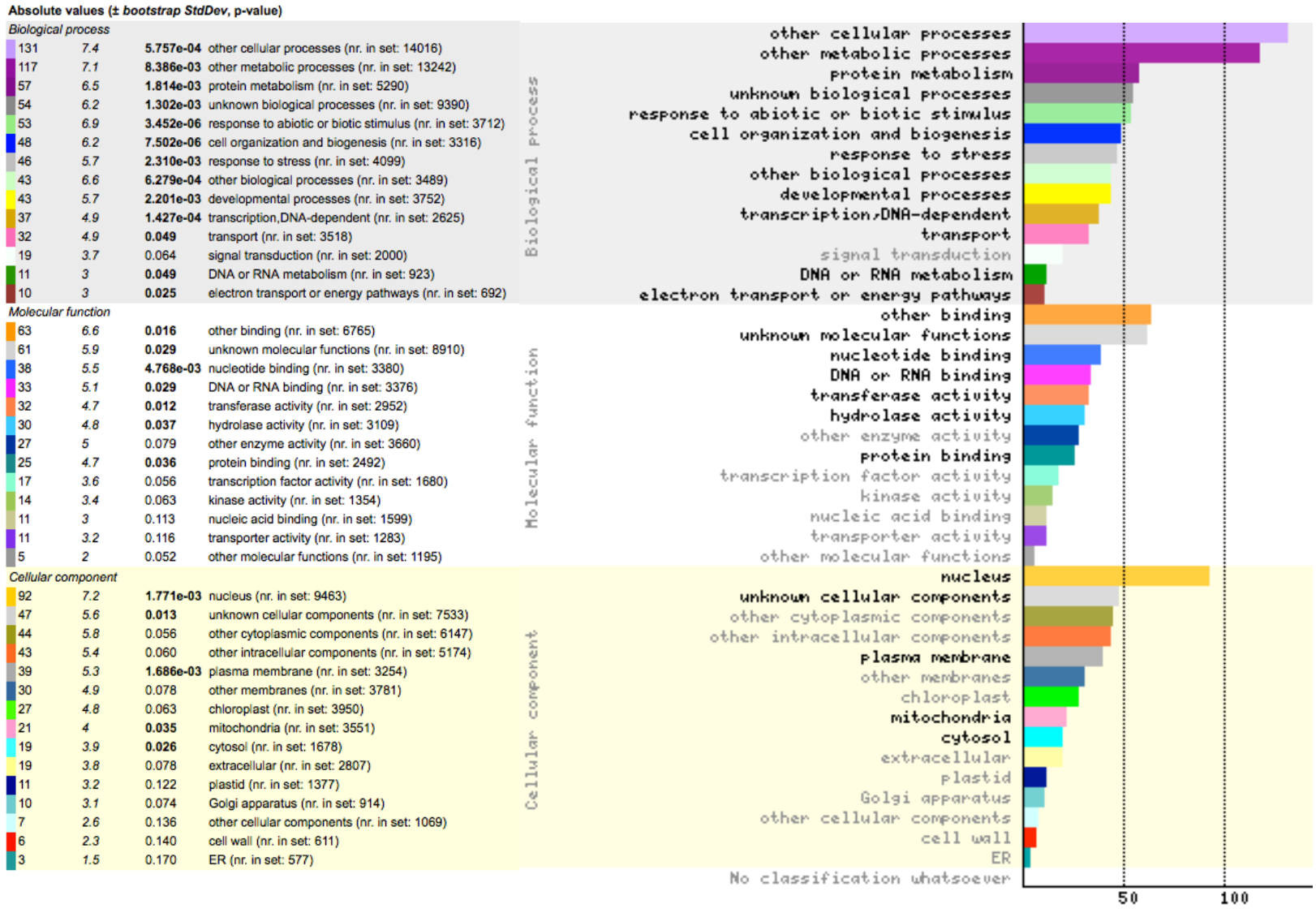
B)

Absolute values (± *bootstrap StdDev*, p-value)

Biological process			
116	7.4	**0.015**	other cellular processes (nr. in set: 14016)
112	7.6	**9.770e-03**	other metabolic processes (nr. in set: 13242)
63	5.7	**0.036**	unknown biological processes (nr. in set: 9390)
48	6	**0.022**	protein metabolism (nr. in set: 5290)
37	5.9	**0.033**	response to stress (nr. in set: 4099)
37	5.5	**6.407e-03**	other biological processes (nr. in set: 3489)
37	5.6	**0.015**	developmental processes (nr. in set: 3752)
36	5.9	**0.019**	response to abiotic or biotic stimulus (nr. in set: 3712)
35	5.8	**2.640e-04**	transcription,DNA-dependent (nr. in set: 2625)
30	4.5	0.058	transport (nr. in set: 3518)
29	5.6	0.053	cell organization and biogenesis (nr. in set: 3316)
18	3.7	0.070	signal transduction (nr. in set: 2000)
8	2.7	0.071	electron transport or energy pathways (nr. in set: 692)
6	2.3	0.154	DNA or RNA metabolism (nr. in set: 923)
Molecular function			
57	6	**0.035**	other binding (nr. in set: 6765)
53	6	**8.219e-03**	unknown molecular functions (nr. in set: 8910)
34	5.8	**0.015**	DNA or RNA binding (nr. in set: 3376)
29	4.7	0.058	nucleotide binding (nr. in set: 3380)
28	5	**0.034**	transferase activity (nr. in set: 2952)
26	5	0.069	hydrolase activity (nr. in set: 3109)
26	4.7	**1.945e-04**	transcription factor activity (nr. in set: 1680)
25	4.5	0.076	other enzyme activity (nr. in set: 3660)
21	4.4	0.077	protein binding (nr. in set: 2492)
17	4	**0.012**	kinase activity (nr. in set: 1354)
11	3	0.096	other molecular functions (nr. in set: 1195)
10	2.6	0.127	transporter activity (nr. in set: 1283)
8	2.4	0.066	nucleic acid binding (nr. in set: 1599)
2	1.3	0.232	receptor binding or activity (nr. in set: 175)
Cellular component			
87	6.5	**3.224e-03**	nucleus (nr. in set: 9463)
51	5.4	**0.047**	unknown cellular components (nr. in set: 7533)
42	5.7	0.056	other cytoplasmic components (nr. in set: 6147)
29	4.7	0.069	mitochondria (nr. in set: 3551)
27	4	0.069	plasma membrane (nr. in set: 3254)
26	4.5	**5.307e-03**	other intracellular components (nr. in set: 5174)
23	4.1	0.050	other membranes (nr. in set: 3781)
22	4.8	**0.028**	chloroplast (nr. in set: 3950)
18	3.9	0.078	extracellular (nr. in set: 2807)
16	3.9	0.064	cytosol (nr. in set: 1678)
10	2.8	0.100	other cellular components (nr. in set: 1069)
8	3	0.108	plastid (nr. in set: 1377)
6	2.5	0.132	cell wall (nr. in set: 611)
4	1.9	0.098	Golgi apparatus (nr. in set: 914)
3	1.8	0.181	ER (nr. in set: 577)



C)


Absolute values (± bootstrap StdDev, p-value)
Biological process
131 7.4 5.757e-04 other cellular processes (nr. in set: 14016)
117 7.1 8.386e-03 other metabolic processes (nr. in set: 13242)
57 6.5 1.814e-03 protein metabolism (nr. in set: 5290)
54 6.2 1.302e-03 unknown biological processes (nr. in set: 9390)
53 6.9 3.452e-06 response to abiotic or biotic stimulus (nr. in set: 3712)
48 6.2 7.502e-06 cell organization and biogenesis (nr. in set: 3316)
46 5.7 2.310e-03 response to stress (nr. in set: 4099)
43 6.6 6.279e-04 other biological processes (nr. in set: 3489)
43 5.7 2.201e-03 developmental processes (nr. in set: 3752)
37 4.9 1.427e-04 transcription,DNA-dependent (nr. in set: 2625)
32 4.9 0.049 transport (nr. in set: 3518)
19 3.7 0.064 signal transduction (nr. in set: 2000)
11 3 0.049 DNA or RNA metabolism (nr. in set: 923)
10 3 0.025 electron transport or energy pathways (nr. in set: 692)
Molecular function
63 6.6 0.016 other binding (nr. in set: 6765)
61 5.9 0.029 unknown molecular functions (nr. in set: 8910)
38 5.5 4.768e-03 nucleotide binding (nr. in set: 3380)
33 5.1 0.029 DNA or RNA binding (nr. in set: 3376)
32 4.7 0.012 transferase activity (nr. in set: 2952)
30 4.8 0.037 hydrolase activity (nr. in set: 3109)
27 5 0.079 other enzyme activity (nr. in set: 3660)
25 4.7 0.036 protein binding (nr. in set: 2492)
17 3.6 0.056 transcription factor activity (nr. in set: 1680)
14 3.4 0.063 kinase activity (nr. in set: 1354)
11 3 0.113 nucleic acid binding (nr. in set: 1599)
11 3.2 0.116 transporter activity (nr. in set: 1283)
5 2 0.052 other molecular functions (nr. in set: 1195)
Cellular component
92 7.2 1.771e-03 nucleus (nr. in set: 9463)
47 5.6 0.013 unknown cellular components (nr. in set: 7533)
44 5.8 0.056 other cytoplasmic components (nr. in set: 6147)
43 5.4 0.060 other intracellular components (nr. in set: 5174)
39 5.3 1.686e-03 plasma membrane (nr. in set: 3254)
30 4.9 0.078 other membranes (nr. in set: 3781)
27 4.8 0.063 chloroplast (nr. in set: 3950)
21 4 0.035 mitochondria (nr. in set: 3551)
19 3.9 0.026 cytosol (nr. in set: 1678)
19 3.8 0.078 extracellular (nr. in set: 2807)
11 3.2 0.122 plastid (nr. in set: 1377)
10 3.1 0.074 Golgi apparatus (nr. in set: 914)
7 2.6 0.136 other cellular components (nr. in set: 1069)
6 2.3 0.140 cell wall (nr. in set: 611)
3 1.5 0.170 ER (nr. in set: 577)
Biological process
other cellular processes
other metabolic processes
protein metabolism
unknown biological processes
response to abiotic or biotic stimulus
cell organization and biogenesis
response to stress
other biological processes
developmental processes
transcription,DNA-dependent
transport
signal transduction
DNA or RNA metabolism
electron transport or energy pathways
Molecular function
other binding
unknown molecular functions
nucleotide binding
DNA or RNA binding
transferase activity
hydrolase activity
other enzyme activity
protein binding
transcription factor activity
kinase activity
nucleic acid binding
transporter activity
other molecular functions
Cellular component
nucleus
unknown cellular components
other cytoplasmic components
other intracellular components
plasma membrane
other membranes
chloroplast
mitochondria
cytosol
extracellular
plastid
Golgi apparatus
other cellular components
cell wall
ER
No classification whatsoever
50
100

D)

Absolute values (± *bootstrap StdDev*, p-value)

Biological process			
48	*4.8*	0.067	other cellular processes (nr. in set: 14016)
44	*4.8*	0.057	other metabolic processes (nr. in set: 13242)
27	*4.3*	**0.028**	unknown biological processes (nr. in set: 9390)
20	*3.7*	**0.020**	response to abiotic or biotic stimulus (nr. in set: 3712)
19	*3.6*	0.054	response to stress (nr. in set: 4099)
19	*3.6*	**0.033**	developmental processes (nr. in set: 3752)
18	*3.8*	**0.033**	other biological processes (nr. in set: 3489)
18	*3.7*	0.098	protein metabolism (nr. in set: 5290)
14	*3.4*	**7.663e-03**	signal transduction (nr. in set: 2000)
12	*3.2*	0.122	cell organization and biogenesis (nr. in set: 3316)
10	*2.8*	0.091	transport (nr. in set: 3518)
10	*2.7*	0.130	transcription,DNA-dependent (nr. in set: 2625)
2	*1*	0.258	electron transport or energy pathways (nr. in set: 692)
1	*0.6*	0.114	DNA or RNA metabolism (nr. in set: 923)
Molecular function			
31	*3.9*	**0.030**	other binding (nr. in set: 6765)
23	*3.6*	**0.011**	unknown molecular functions (nr. in set: 8910)
18	*3.5*	**0.015**	hydrolase activity (nr. in set: 3109)
17	*3.7*	0.061	other enzyme activity (nr. in set: 3660)
16	*3.3*	0.060	nucleotide binding (nr. in set: 3380)
14	*3.4*	0.068	transferase activity (nr. in set: 2952)
12	*2.7*	0.075	protein binding (nr. in set: 2492)
11	*2.7*	0.117	DNA or RNA binding (nr. in set: 3376)
8	*2.5*	0.109	transcription factor activity (nr. in set: 1680)
6	*2.5*	0.123	other molecular functions (nr. in set: 1195)
6	*2.1*	0.165	nucleic acid binding (nr. in set: 1599)
6	*1.9*	0.147	kinase activity (nr. in set: 1354)
5	*1.9*	0.178	transporter activity (nr. in set: 1283)
2	*1.1*	0.107	receptor binding or activity (nr. in set: 175)
Cellular component			
31	*5*	0.066	nucleus (nr. in set: 9463)
24	*4.4*	0.069	unknown cellular components (nr. in set: 7533)
18	*3.3*	**0.036**	mitochondria (nr. in set: 3551)
14	*3.3*	**0.012**	other cytoplasmic components (nr. in set: 6147)
14	*3.3*	0.092	plasma membrane (nr. in set: 3254)
14	*3.2*	0.051	other intracellular components (nr. in set: 5174)
12	*3.1*	0.103	extracellular (nr. in set: 2807)
9	*2.7*	**0.048**	other membranes (nr. in set: 3781)
6	*2.4*	0.067	Golgi apparatus (nr. in set: 914)
5	*2.1*	0.161	cytosol (nr. in set: 1678)
4	*1.8*	**5.815e-04**	chloroplast (nr. in set: 3950)
4	*1.7*	0.110	cell wall (nr. in set: 611)
4	*1.7*	0.199	other cellular components (nr. in set: 1069)
1	*0.8*	**0.031**	plastid (nr. in set: 1377)
1	*0.6*	0.257	ER (nr. in set: 577)

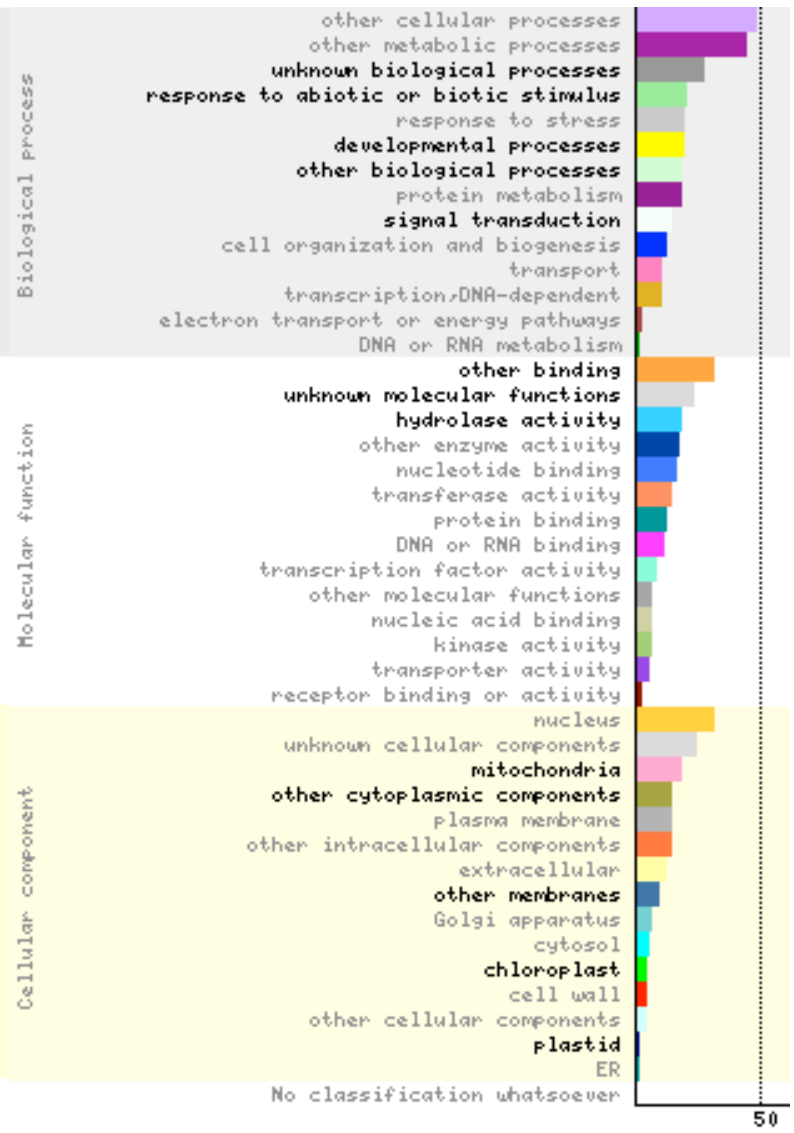


Figure 26 Cellular-level functions of genes exhibiting signatures of selection Genes closest to loci exhibiting signatures of selection (see Chapter **4.1.2**) in samples taken from a single habitat type were analysed using the Classification Superviewer tool (Provart & Zhu 2003) from the Bio-Analytic Resource toolset (http://bar.utoronto.ca/). When given a list of *A. thaliana* genes, this tool compares the frequencies of classes of Gene Ontology terms associated with the input genes to that expected from an equally sized set of genes selected at random. The analysis aimed to identify any functional classifications of the genes bearing signatures of selection which were statistically over-represented compared to expectations from chance, thereby indicating both the probable general driving factors of selection, and productive hypotheses for future experiments. This analysis was repeated across sample groups resident in **A)** wall/rocky outcrop, **B)** garden, **C)** other/unnamed habitats, and from **D)** all UK habitats combined. Railway-type habitats were excluded from this analysis since the small number of samples prevented the meaningful identification of loci under selection. Output of the program is arranged into two formats. The table on the left side lists functional annotation classes, and for each, the number of genes carrying that functional annotation, the standard deviation in the number of genes carrying that functional classification from the expected mean, and the p-value showing the likelihood of the supplied set of genes containing the observed number of functional annotations of that type. The bar chart on the right shows the number of genes in the input set carrying each functional classification. Functional classifications in both plots are sorted by the number of annotations for that class, in descending order. Note that genes may carry several functional classifications, so the total number of classifications exceeds the number of input genes. Functional classifications typed in bold are significantly over-represented in the input set of genes, compared to chance expectations. Classifications of response to abiotic/biotic stimuli and developmental processes are significantly over-represented across all habitats. Classifications of response to stress are also significantly over-represented in individual habitat types. Together, this demonstrates that SelectionFinder analysis is capable of successfully identifying loci containing genotypic variation responsible for adaptation to environmental factors.

that are significantly over-represented at haplotype centre-points across all habitats compared to chance expectations, and may explain selective responses of UK populations to external abiotic and/or biotic factors. Genes associated with stress responses are also over-represented within different habitats, but not across the UK population as whole, suggesting that factors driving selection upon stress response phenotypes may be specific to particular habitats (*i.e.,* local adaptation) rather than uniform across the UK.

Lists of genes falling within haplotypes marked as being favoured by selection in each of the analysed habitat types, along with their functional annotations, are shown in Appendix **2**; the samples actually possessing said haplotypes are shown in Appendix **3**. A list of LRR-type genes that may be undergoing positive selection relative to different habitats is particularly interesting (Appendix **6.1**). The distribution across the genome of these haplotypes, plotted against the rate of meiotic crossovers, is shown for each habitat type in Figure **27**.

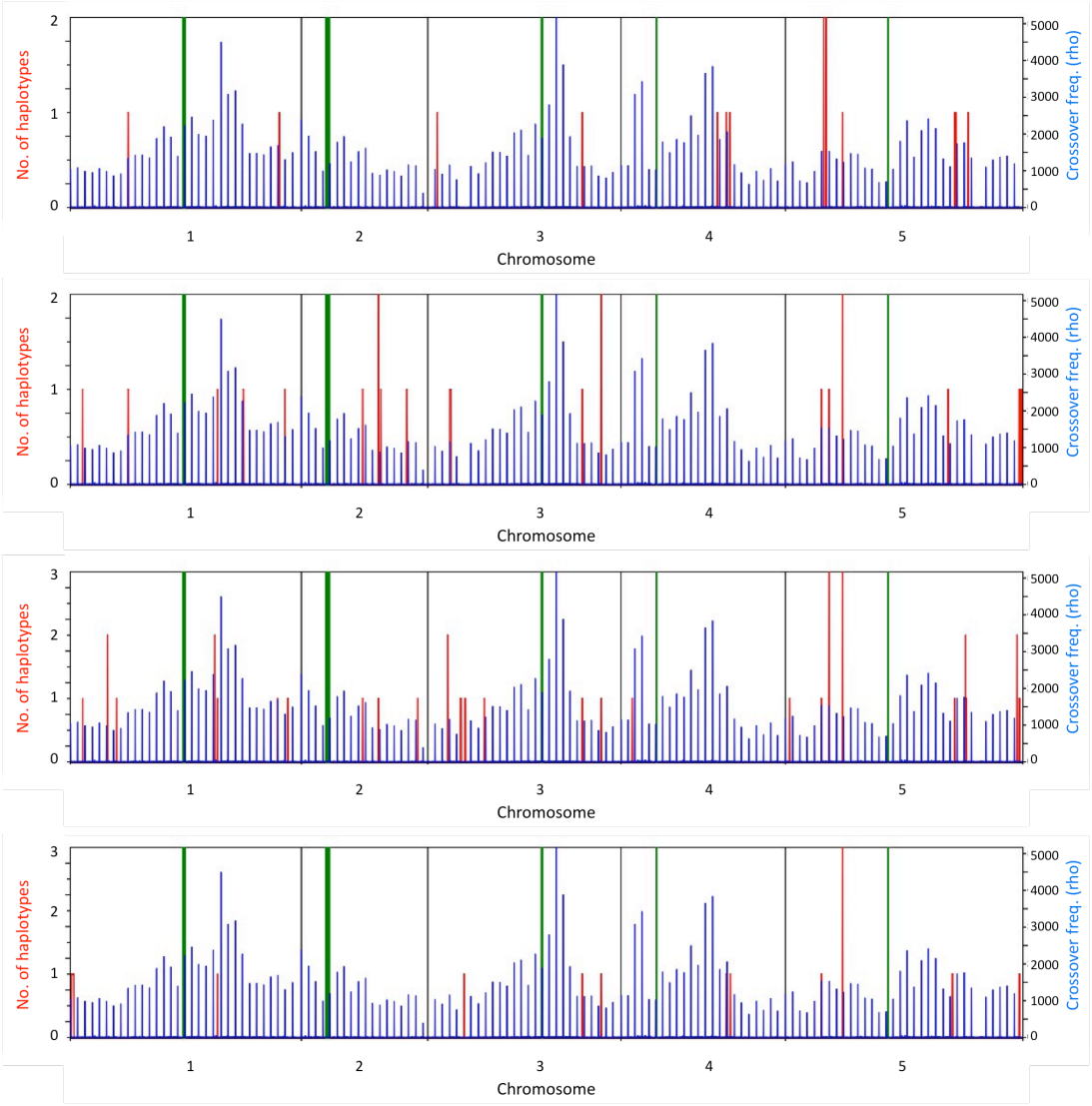


Figure 27 Loci of haplotypes found by SelectionFinder to be under selection plotted against crossover rate showing the number of haplotypes (**RED** series) plotted alongside the rate of meiotic crossovers at corresponding loci (**BLUE** series). From top to bottom are plots showing the numbers of loci under selection in the Wall/Outcrop, Garden, Other and UK-wide habitat types. No significant relationship was found between crossover rate and number of haplotypes possessing signatures of selection within any of the habitat types.

Quantifications of crossover rate at each LRR-type gene and flowering time-associated gene found to show signatures of selection are also presented in Appendices **6.1** and **6.2**, respectively. None of these genes appears to be located near to loci undergoing a high frequency of recombination.

The UK-wide Ka/Ks ratio (non-functional/functional mutation ratio) of genes in this list was estimated using summary data from the Arabidopsis 1001 Genomes database (*http://signal.salk.edu/atg1001/3.0/gebrowser.php*). Several

of these genes show a higher proportion of non-synonymous mutations, further indicating the effect of selection pressure. Interestingly, no genotype specific *R* genes (*i.e.,* conferring resistance to pathogens that are used widely for laboratory research such as *Pseudomonas*, *Hyaloperonospora*, *Albugo* or *Leptosphaeria*) were identified in this analysis.

A similar analysis provides a list of genes that may indicate selection affecting flowering time (Appendix **6.2**), using candidate genes from a GWAS analysis by Ehrenreich *et al.* (Ehrenreich et al. 2009). Six genes are marked as candidates for selection, five of which show further evidence of selection through the Ka/Ks ratio.

4.4 DISCUSSION

4.4.1 EVALUATION OF SELECTIONFINDER

The SelectionFinder tool provides a means of interrogating large DNA sequence datasets with predictive modeling to identify regions in a genome that may have been affected by selective pressure in local populations. The tool is flexible for use with species of different life styles. For example, in addition to the control over dispersal and recombination already described, the tool also allows for control of the number of seeds produced by a given individual, and for the introduction of variation in this characteristic between individuals. Indeed, so long as the number of seeds produced per plant is set such that more offspring are produced than are allowed to survive to produce the next generation, and certain genotypes are set to cause the simulated organisms possessing them to be more likely to survive and themselves reproduce, this simulated population possesses all four of the requirements laid out by Darwin in order for evolution by natural selection to occur.

In other research contexts, this tool may find application for the study of hypothetical evolutionary scenarios in which population structure plays a role in determining the course of a species' evolution. Control over the proportion of seeds produced per generation relative to those actually recruited to the subsequent generation may be employed to fit the tool to species following very different lifestyles (for example, herbaceous annuals that rely on the production of large numbers of seeds for adequate propagation), or to reflect more subtle differences in lifestyle within a species or between closely related species such as *Arabidopsis thaliana* and *Arabidopsis lyrata*. An analysis of this type may comprise of comparisons between simulated sets of genotypes in a similar manner to the comparison between populations resident in the habitat types represented in the data available to this project.

Reliably drawing conclusions from SelectionFinder analysis relies upon an accurate simulation of the gene flow of the wild population, however, in order

that genotypic variation that is subject to the influence of selective pressures can be distinguished from selectively neutral variation. Verification of the population model used by PopAger and SelectionFinder shows that the simulation approximates the general properties of the wild population across the UK (see Chapter **2.3.5**) and at more local scales (see Chapter **3.3.6**).

In the context of SelectionFinder, though, it may sometimes be helpful to explicitly account for known population genetic factors that differ from the model by which the population as a whole is represented. For example, certain *A. thaliana* accessions are known to be adapted to low-disturbance wall sites, and to produce an atypically small number of seeds (Chattaway, *unpublished work*). In theory, this could be explicitly accounted for by SelectionFinder; however, in practice – at least in this case – there are reasons to avoid doing so.

A population model may be constructed for each habitat type and run independently; however, this would fail to account for gene flow from other habitats, thus negating the point of running the simulation. Additionally, the small sample size within each habitat type would leave SelectionFinder with a limited ability to detect haplotypes under selection, since stochastic effects would limit the resolution of the threshold at which SelectionFinder would mark haplotypes as possessing signatures of selection.

Alternatively, new model parameters may be dynamically specified for individuals possessing haplotypes known to influence the parameters in question for individuals at specific sites. However, this risks introducing further biases to the model unless the strength and circumstances of selection are accurately known beforehand.

Since SelectionFinder operates on the principle of detecting significant departures from the population-wide average of the distributions of neutral variation, and since the demographic model on which the simulated population is based provides a suitable approximation of the wild population, the best

solution – at least, in the case of this project – is to accept any known cases of selection as positive controls. Indeed, the results of SelectionFinder analysis do indicate that different habitat types may possess different optimum points along the r/K spectrum, and that selection is acting upon the *A. thaliana* population in this respect.

Originally, this project aimed to detect balancing selection as well as positive selection. However, this has proved impossible, due to the tendency for the SNP data available to this project to approach or reach maximum possible diversity across even the relatively small numbers of sampled accessions within each habitat type. Detection of balancing selection may be achieved in the future, however, if the same approach is applied to yet more dense data, such as resequencing data, in which regions of uncommonly high allelic diversity are more apparent. Indeed, given the potential for haplotype-based methods to recognise partial or soft sweeps as well as full sweeps that reach fixation, the method outlined in this project remains a promising means of detecting balancing selection and of making comparisons with positive selection.

Additionally, plans were made to further extend the SelectionFinder analysis to incorporate spatial data to a much greater degree – i.e., to test for haplotypes showing an unusual degree of spatial dispersal throughout the population for their lengths or their frequencies – but due to time constraints, this analysis was not carried out. This remains a potentially useful avenue of future research, which future investigators may wish to follow up.

Although the SelectionFinder method is highly resistant to bias arising from selection of polymorphisms, it is less resistant to biased selection of sample accessions. Genotype data utilised in this project was selected in order to capture as much of the total allelic diversity of the global *A. thaliana* population as reasonably possible, rather than to accurately assess the frequencies of alleles within any given local area. While this is likely to affect the robustness of specific ecological conclusions drawn from this type of analysis – such as the

frequencies of haplotypes currently undergoing selection within a given area, for example – this is unlikely to significantly affect more basic conclusions, such as which loci selection is actually acting upon. More precisely, this may somewhat confound efforts to detect balancing selection through heightened genetic variability in surrounding loci, as indeed occurred in this chapter; but conversely, this also eases the detection of selective sweeps (which, in any case, may still occur under the unstable equilibrium of balancing selection, though are likely to remain only 'partial' sweeps). Future research along these lines may opt to follow sampling procedures, or use datasets, that better reflect the real allelic frequencies of a population in order to remove these factors from consideration.

4.4.2 ECOLOGICAL CONCLUSIONS FROM SELECTIONFINDER

SelectionFinder analysis revealed an over-representation of haplotypes that contain genes associated with responses to biotic and abiotic factors, indicating the predictive power of the application for investigating adaptation of populations to local environmental conditions, and most importantly, for informing the design of field experiments to test predictions. For example, the signatures of selection at known *R* genes and other defense-related genes indicate that *A. thaliana* populations in the UK have been affected by selection pressure from microbial activity.

More than sixty genes containing a LRR domain were found as potential candidates for selection. Of those, at least 12 have previously been reported to be in some way involved in the regulation of induced cell death (the hypersensitive response) and/or other responses to microbial infection. While some of the genes in this class showed evidence of sweeps across several habitat types, the majority showed evidence of sweeps in a single habitat type. This may suggest that local adaptation to specific habitat types occurs across the UK, and in turn suggests significant differences in both the set of pathogenic

species and the environmental conditions encountered across different habitats.

A number of the LRR genes found to possess signatures of selection control aspects of the phenotype relating to the abiotic conditions of the environment. For example, 3 genes - AT1G05700, AT1G17610 and AT3G20600 (see (Kreps et al. 2002; Wang et al. 2013)) – have previously been reported to confer tolerance to cold conditions; intriguingly, within each habitat type, only one of these three genes exhibits signatures of selection. This may constitute an example of local adaptation.

A signature of selection corresponding to the gene AT2G26290 was found in the population in the wall/rocky outcrop habitat. The kinase produced by this gene becomes activated in conditions of dehydration- or salt-stress. Since wall/outcrop habitats are likely to retain less water than other habitats (due to the relative lack of soil), it is reasonable to propose that selection pressures more strongly favour alleles conferring an improved tolerance to dehydration in the wall/outcrop habitat than in other habitats.

Several signatures of selection were discovered at loci associated with flowering time, particularly within the garden habitat type (see Appendix **6.2**), and also at a number of LRR loci associated with growth and development – for example, AT1G75820 (associated with root development (Stahl et al. 2013)), AT3G49670 (associated with anther and meristem development (DeYoung et al. 2006; Hord et al. 2006)) and AT4G20270 (also associated with meristem development and cold tolerance (DeYoung et al. 2006)). The Ka/Ks ratios at all but one of the flowering time-associated loci indicate strong positive selection. In conjunction with the conclusions regarding optimum points on the r/K spectrum discussed below, this may suggest the presence of unique selection pressures for variation in life history traits within this habitat type.

A fourth group of selection signatures corresponds to a set of several genes of the DAR family - AT5G66610, AT5G66620, AT5G66630 and AT5G66640 - which play a role in the control of organ and seed size (Li et al. 2008). The gene AT2G39830 also exhibits a similar signature of selection, and is also a determinant of seed size (Li et al. 2008). This signature of selection does not appear across all habitats. Larger seeds require parents to dedicate a greater amount of resources towards the growth of their offspring, which also determines the number of offspring any given individual is ultimately capable of producing. This may, therefore, indicate that different habitats favour different optimal points in the trade-off between investment of resources in offspring and in growth, and consequently different points along the r/K strategy continuum. The degree of habitat variability is known to affect the optimal point of this trade-off; experimentation has shown that a greater degree of environmental variability favours a strategy of higher fecundity, necessitating a shift in investment of resources away from the growth and longevity of the parent and towards investment in larger numbers of offspring (Rose & Charlesworth 1980).

Interestingly, the signature of selection for the *DAR* genes was detected in garden habitats (which undergo a substantial degree of human-caused disturbance) but not in low-disturbance 'wall/outcrop' habitats. A hypothetical explanation may be that *A. thaliana* populations sampled from garden habitats have been undergoing selection at loci affecting seed production in response to factors associated with habitat disturbance (*e.g.,* release of nutrients from cultivation), whereas populations that were sampled from wall/outcrop habitats are fully adapted for survival in the harsh, low nutrient conditions and consequently genetically uniform at the same loci.

This variation in selection for r/K strategies may explain the apparent migratory history of the UK population inferred from PCA clustering of genotypes in Chapter **3.3.1**. Certain accessions in low-disturbance wall/outcrop habitats are known to set

smaller numbers of seeds than the population-wide average (see Chapter **4.4.1**). Since human disturbance of the environment at these sites is minimal, a population that establishes at such sites could have adapted to become highly specialised to the habitat type represented by walls and rocky outcrops (and, thus, also somewhat genotypically differentiated from mainland populations). As humans also created other, more disturbed habitats – represented by gardens – in more recent times, there would therefore exist selection pressures in these habitats towards an r strategy, but the populations in less disturbed habitats would remain unaffected, and would therefore not show any signatures of selection due to their already high degree of adaptation to those habitats. This plausible scenario may be tested by common garden experiments (see below. If supported by further evidence, this demonstrates that *A. thaliana* is a suitable case study for the long-term effects of human actions on the selection pressures exerted on wild populations by their environments.

Researchers wishing to further investigate instances of local adaptation should design common garden experiments using pools of samples drawn from all relevant habitat types. For example, an experiment to further investigate the potential multiple instances of local adaptation in favour of cold tolerance could involve sowing seeds collected from representative populations within each habitat type, at a set of common gardens across all 3 habitat types. Genotypic assays of alleles at the 3 listed loci taken over the course of several generations would then reveal which alleles are favoured in each circumstance, and would also more clearly show the type of selection occurring (i.e., balancing, directional, stabilising, etc.); further hypotheses may then be proposed as to why certain alleles are favoured in a given situation.

4.4.3 PLANT-PATHOGEN INTERACTION CONCLUSIONS FROM SELECTIONFINDER

Many of the LRRs found to possess signatures of selection have been reported to be involved in defence against pathogens. A significant range of pathogens are

represented via their *R* genes or PRRs within the loci marked by SelectionFinder, spanning bacteria (AT1G55020 – *Xanthomonas campestris campestris* resistance (Montillet et al. 2013); AT3G20600 – component of systemic acquired resistance to many bacteria, incl. *Pseudomonas syringae* (Lewis et al. 2010)), viruses (AT1G05760 – tobacco etch virus resistance (Cosson & Sofer 2010); AT5G16000 – antiviral signalling (Sakamoto et al. 2012)), nematodes (AT1G75820 – detection of nematode effectors (Replogle et al. 2013)), fungi (AT1G71830 – resistance against *Verticillium* spp. (Fradin et al. 2011); AT1G72300 – resistance to *Alternaria brassicicola* (Mosher et al. 2013)) and oomycetes (AT4G20380 – resistance to *Hyaloperonospora arabidopsidis* (Cooper et al. 2008)). Two genes – AT1G74360 and AT3G14840 – are also triggered by the detection of oviposition by butterflies of the Pieridae family (Little et al. 2007). A number of genes involved in regulating the hypersensitive response are also present.

One candidate TIR-NBS-LRR gene described by Kim *et al.* (Kim et al. 2012) and named *VICTR* (AT5G46520) is of particular interest. This gene encodes a receptor protein that responds to treatment with a small signaling molecule DPFM (5-(3,4-dichlorophenyl)furan-2-yl]-piperidine-1-ylmethanethione), and causes a localised arrest of primary growth in the root meristem upon detection of that compound.

Activation of defence pathways and restriction of root growth is likely to limit the potential damage caused to a plant encountering soil-borne pathogens, and could therefore be a significant determinant of fitness. The specific response of a local cessation of root growth (rather than hypersensitive cell death) may also be adaptive, since programmed death of root cells may have a greater detrimental effect on the ability of the plant to flourish in adulthood than attack by pathogens. Alternatively, the cessation of growth may simply limit further exposure to pathogen attack though the simple expedient of avoiding placing vulnerable tissues in areas found to contain pathogens. As such, it may be that the response mediated by *VICTR* can be regarded as an optimal point in an

evolutionary trade-off. Interestingly, *VICTR* was also noted to share a high degree of homology with other genes known to confer pathogen-specific resistance, including the *R* gene *RPS6* (which detects the presence of the *hopA1* effector from *P. syringae* pv. *syringae* (Kim et al. 2009)) and the white rust resistance gene *WRR1* (previously *RAC1*) (Kim et al. 2009; Borhan et al. 2004).

An important membrane-bound receptor-like gene named *EFR* (AT5G20480) (Zipfel et al. 2006) was also identified in the SelectionFinder analysis. This gene encodes a homolog of the PRR kinase *FLS2*, which confers recognition of a pathogen-associated molecular pattern (PAMP) in bacterial flagellin. Upon detection of the EF-Tu PAMP produced by the pathogenic bacterium *Agrobacterium tumefaciens, EFR* induces a similar response to that induced by *FLS2* (Zipfel et al. 2006).

The lack of detection for positive selection in genes found to be associated with either total or partial white rust resistance (*WRR1*/*RAC1*, *WRR4*, *WRR5* and *WRR6*) by MAGIC mapping (see Chapter **4.3.1**) suggests that *Albugo candida* is not imposing a significant selection pressure in the temporal and spatial scales of *A. thaliana* populations that were sampled in this study. However, the possibility cannot yet be ruled out that white rust resistance instead follows the pattern of small, quantitative changes across a large number of genes leading to small overall changes in phenotype. For example, *DAR5* is one member of the family that was identified by SelectionFinder and may play a role in response to *Albugo candida*. This gene is located in a locus designated *WRR7* which is associated with a weak 'loss-of-turgidity' response that permits colonization of tissue by *A. candida* but impedes asexual reproduction (Taylor, Cevik and Holub, unpublished). Another locus - AT4G20380 - was reported in previous experiments (Cooper et al. 2008) to be suppressed by Albugo infection, leaving the host vulnerable to extensive infection by other biotrophic pathogens, such as *Hyaloperonospora arabidopsis*, that would normally be halted by the hypersensitive response regulated by this gene's protein product. Along with

further findings discussed below, this underscores that a complete picture of plant-pathogen interactions must come from viewing the whole system as a complex network of interactions, as proposed in Chapter **1.7.1**.

Most notably, some 18 of the 62 loci marked by SelectionFinder are part of a complex web of interactions that was found to be activated in response to infection by geminivirus infection (Ascencio-Ibáñez et al. 2008). Many of these genes are also known to be associated other functions, including responses to stress and development. In order to gain a complete understanding of the interactions between plants and pathogens, and to understand the outcomes of pathogen attacks in the real world and the ultimate success or failure of genotypes in the face of these attacks, it is now clear that we need to examine more than direct interactions between pathogen effectors and host *R* genes. It is necessary to view these interactions in the wider context of the entire molecular machinery, developmental processes and ecological circumstances of the host and pathogen.

Given the relative adaptability of this approach to next-generation resequencing data such as that produced by the 1001 Genomes Project, future analyses along the lines laid out in this chapter may be employed to answer a very broad range of questions relating to ecology and evolution. Whole-genome analyses of the described in this chapter remain one possibility; however, the approach may also be used simply to investigate the possibility of selection acting upon specific loci already suspected, through other work, to be subject to selection.

CHAPTER 5: OVERALL DISCUSSION

Through the development and application of the SelectionFinder analysis tool (Chapter **4**), this project has demonstrated not only that the population structure does not interfere with ecological genetic and evolutionary analyses, but that it can actually be turned to our advantage. A very simple model of this population structure, arising from a simple exponential decay in dispersal likelihood over distance, gives a robust set of expectations against which observations of both the spatial distributions and population frequencies may be compared.

This project set out to resolve several issues relating to the ecology and evolution of *A. thaliana* as a model plant species, particularly in regard to the use of the UK population as a natural evolutionary experiment.

Analysis of population structure confirmed the general validity of the isolation by distance model in explaining the observed distribution of genotypes across the British Isles, and revealed a degree of emergence of this structure indicating a clearly more recent establishment of the UK population than those on the European mainland. Another question arises from this analysis, however, since it is unclear why *A. thaliana* should have become established in the UK for a shorter amount of time than other habitats at comparable latitudes that opened up at the end of the Pleistocene, given that the current British Isles are known to have been connected to the mainland via a land bridge (Scourse & Austin 1995).

Clustering analysis was used to probe this matter, and found evidence of several distinct migration events. Further analysis failed to determine the amount of time since the founding of the population, but demonstrated the validity of a model based on scaled exponential decay of seed dispersal likelihood over distance.

Given this information, a simulation-based approach may yet yield a sound estimate of the time at which the UK population became established, though further analysis of the precise sources and migration events into the UK may also enable a tentative estimate based on coalescent time.

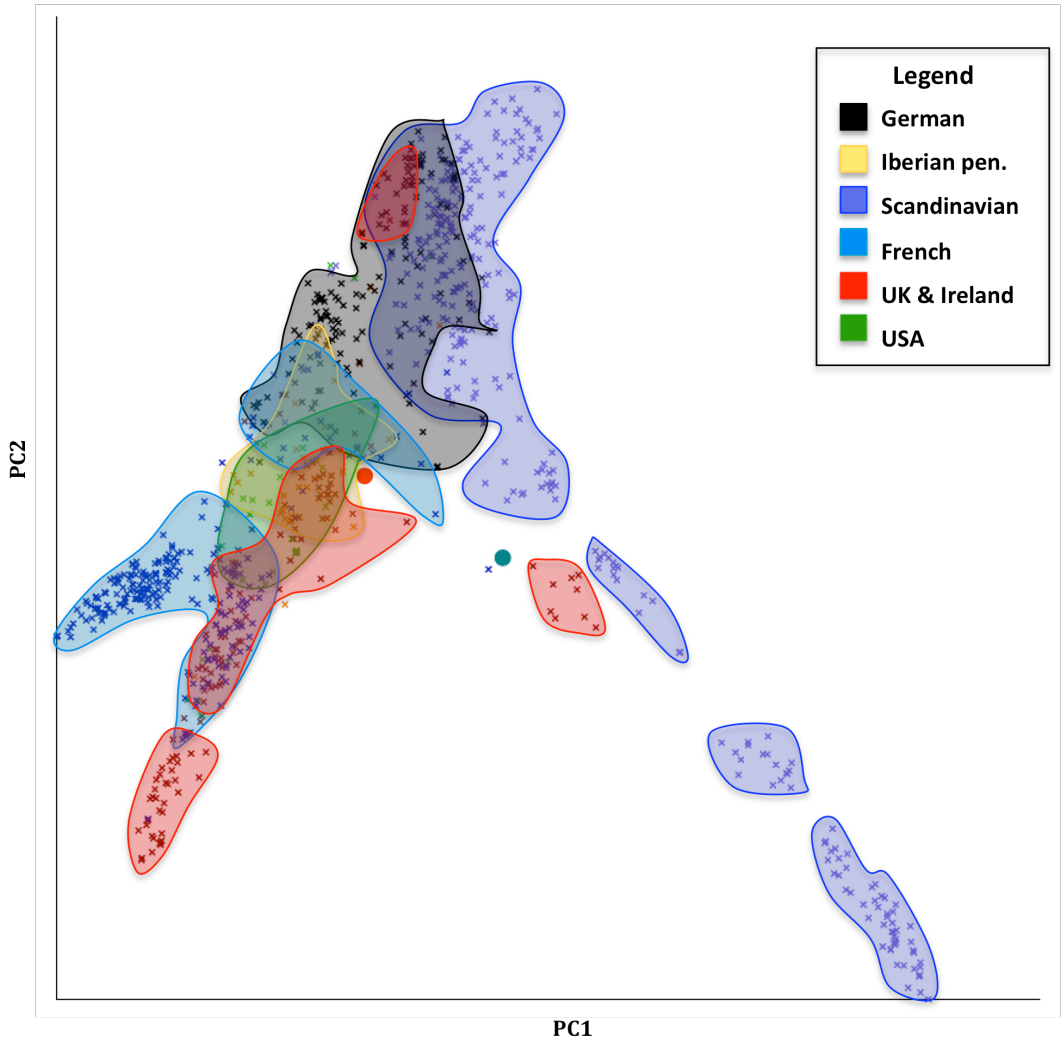


Figure 28 Principal coordinate analysis of European genotypes with accessions Bur-0 (BLUE spot) and Cvi-0 (ORANGE spot) marked. Bur-0 clusters near to the UK-Scandinavian cluster identified in Chapter 3, while Cvi-0 clusters near to the UK-Iberian-French cluster.

When this knowledge was applied to the detection of signatures of selection, a large number of LRR receptor-like genes were found to show evidence of adaptation to some or all of the habitats across the UK. Some of the genes thus marked suggest further field experiments to verify and expand upon the conclusions drawn from the analysis. One of these is inspired by unpublished

work by fellow DTG student Richard Chattaway (*pers. comm.*). This work showed a distinction between two life history phenotypes: a 'nomadic' type, possessing a phenotype of high fecundity and adaptation to a relatively broad range of habitat types; and a 'wall-based' type, characterised by a phenotype of low fecundity and adaptation to habitats in which disturbance is very rare. Two accessions representative of each type – Cvi-0 and Bur-0 respectively – were identified. These two accessions are marked in Figure **28**, showing that the two accessions belong to clearly different genotypic clusters.

These two accessions may be used in common garden experiments comparing their relative fitness across several environment types. Seeds of both accessions should be sown across representative garden and wall habitats, and the frequencies of genes identified by SelectionFinder as exhibiting signatures of selection measured after a period of several years. It is a highly plausible hypothesis for this type of experiment that genotypic variation at the *DAR5* locus explains phenotypic variation between accessions, and thus the fitness of particular life history strategies to different habitats.

The most important outcome of this project, though, is the ability to examine all of these possible selection pressures at once, and therefore to begin to see their combined effects in a way that has rarely been available before. Future research may use approaches like those developed in this project to discover and understand the genotypic variation underlying complex quantitative traits – and how variation in these traits contributes towards the overall fitness of organisms through subtle evolutionary trade-offs and balancing acts.

While this project has fulfilled its stated aims – in that it has developed tools to model population structure *in silico* and to separate signatures of selection from that population structure – some potential complications remain in the application of the approaches developed here to species other than *A. thaliana*. The population structure of *A. thaliana* has proved highly conducive to this type of analysis, not only providing a simple theoretical set of expectations, but also

enabling the simulation of a large population with a much smaller number of simulated individuals through the scaling of dispersal and reproduction parameters (see Chapter **3.4.2**). However, populations of other species are unlikely to follow this model unless they are also distributed over a large and essentially continuous geographic range. In cases better represented as several distinct and relatively isolated groups, the methods developed in this project should theoretically provide useful results, though conventional population genetic and ecological analyses would be more appropriate.

Instead of a whole-genome analysis, future work may also seek to more precisely analyse a smaller portion of a genome, such as gene family or cluster, for signatures of selection when testing a specific hypothesis. Resequencing data such as that produced by the 1001 Genomes Project or future projects even broader in scope would be ideal for this type of analysis. Analysis of haplotypes at this high resolution may permit the identification of variation within a gene upon which selection acts, and would certainly also permit the detection of smaller haplotypes, potentially increasing our knowledge of more historically distant selection and demographic events.

Overall, this project presents a bright and exciting future for research at the meeting point of ecology, evolution and genetics, and shows that *Arabidopsis thaliana* still has a key role to play in increasing our understanding of the living world.

REFERENCES

Al-Shehbaz, I.A., Beilstein, M.A. & Kellogg, E.A., 2006. Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant Systematics and Evolution*, 259(2-4), pp.89–120.

Alonso-Blanco, C. & Koornneef, M., 2000. Naturally occurring variation in Arabidopsis: an underexploited resource for plant genetics. *Trends in Plant Science*, 5(1), pp.22–9.

Altschul, S., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), pp.3389–3402.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp.403–10.

Anastasio, A.E., Platt, A., Horton, M., Grotewold, E., Scholl, R., Borevitz, J.O., Nordborg, M. & Bergelson, J., 2011. Source verification of mis-identified Arabidopsis thaliana accessions. *The Plant Journal*, 67(3), pp.554–66.

Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814), pp.796–815.

Arnold, B.C., 1983. *Pareto distributions*, International Co-operative Publishing House.

Arora, V.K., Scinocca, J.F., Boer, G.J., Christian, J.R., Denman, K.L., Flato, G.M., Kharin, V. V., Lee, W.G. & Merryfield, W.J., 2011. Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases. *Geophysical Research Letters*, 38(5).

Ascencio-Ibáñez, J.T., Sozzani, R., Lee, T.-J., Chu, T.-M., Wolfinger, R.D., Cella, R. & Hanley-Bowdoin, L., 2008. Global analysis of Arabidopsis gene expression uncovers a complex array of changes impacting pathogen response and cell cycle during geminivirus infection. *Plant Physiology*, 148(1), pp.436–54.

Astle, W. & Balding, D.J., 2009. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*, 24(4), pp.451–471.

Bakker, E.G., Stahl, E. a, Toomajian, C., Nordborg, M., Kreitman, M. & Bergelson, J., 2006. Distribution of genetic variation within and among local populations of Arabidopsis thaliana over its species range. *Molecular Ecology*, 15(5), pp.1405–18.

Bancroft, I., 2000. Insights into the structural and functional evolution of plant genomes afforded by the nucleotide sequences of chromosomes 2 and 4 of Arabidopsis thaliana. *Yeast*, 17(1), pp.1–5.

Barrett, L.G., Kniskern, J.M., Bodenhausen, N., Zhang, W. & Bergelson, J., 2009. Continua of specificity and virulence in plant host-pathogen interactions: causes and consequences. *The New Phytologist*, 183(3), pp.513–29.

Bateson, W., Waunders, E.R. & Punnett, R.C., 1909. Experimental studies in the physiology of heredity. *Zeitschrift für Induktive Abstammungs- und Vererbungslehre*, 2(1), pp.17–19.

Beck, J.B., Schmuths, H. & Schaal, B. a, 2008. Native range genetic variation in Arabidopsis thaliana is strongly geographically structured and reflects Pleistocene glacial dynamics. *Molecular Ecology*, 17(3), pp.902–15.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. & Haussler, D., 2004. Ultraconserved elements in the human genome. *Science*, 304(5675), pp.1321–5.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), pp.53–9.

Bergelson, J. & Roux, F., 2010. Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nature Reviews Genetics*, 11(12), pp.867–79.

Bergelson, J., Stahl, E., Dudek, S. & Kreitman, M., 1998. Genetic variation within and among populations of Arabidopsis thaliana. *Genetics*, 148(3), pp.1311–23.

Bittner-Eddy, P.D., Crute, I.R., Holub, E.B. & Beynon, J.L., 2000. RPP13 is a simple locus in Arabidopsis thaliana for alleles that specify downy mildew resistance to different avirulence determinants in Peronospora parasitica. *The Plant Journal*, 21(2), pp.177–88.

Blattner, F.R., 1997. The Complete Genome Sequence of Escherichia coli K-12. *Science*, 277(5331), pp.1453–1462.

Blossey, B. & Notzold, R., 1995. Evolution of Increased Competitive Ability in Invasive Nonindigenous Plants: A Hypothesis. *The Journal of Ecology*, 83(5), p.887.

Bomblies, K., Yant, L., Laitinen, R. a, Kim, S.-T., Hollister, J.D., Warthmann, N., Fitz, J. & Weigel, D., 2010. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. *PLoS Genetics*, 6(3), p.e1000890.

Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.-S., Zhu, T., Weigel, D., Berry, C.C., Winzeler, E. & Chory, J., 2003. Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Research*, 13(3), pp.513–23.

Borhan, M.H., Brose, E., Beynon, J.L. & Holub, E.B., 2001. White rust (Albugo candida) resistance loci on three Arabidopsis chromosomes are closely linked to downy mildew (Peronospora parasitica) resistance loci. *Molecular Plant Pathology*, 2(2), pp.87–95.

Borhan, M.H., Gunn, N., Cooper, A., Gulden, S., Tör, M., Rimmer, S.R. & Holub, E.B., 2008. WRR4 encodes a TIR-NB-LRR protein that confers broad-spectrum white rust resistance in Arabidopsis thaliana to four physiological races of Albugo candida. *Molecular Plant-Microbe Interactions*, 21(6), pp.757–68.

Borhan, M.H., Holub, E.B., Beynon, J.L., Rozwadowski, K. & Rimmer, S.R., 2004. The Arabidopsis TIR-NB-LRR gene RAC1 confers resistance to Albugo candida (white rust) and is dependent on EDS1 but not PAD4. *Molecular Plant-Microbe Interactions*, 17(7), pp.711–9.

Boyes, D. & Zayed, A., 2001. Growth stage–based phenotypic analysis of Arabidopsis a model for high throughput functional genomics in plants. *The Plant Cell*, 13(7), pp.1499–510.

Boyes, D.C., Nam, J. & Dangl, J.L., 1998. The Arabidopsis thaliana RPM1 disease resistance gene product is a peripheral plasma membrane protein that is degraded coincident with the hypersensitive response. *Proceedings of the National Academy of Sciences*, 95(26), pp.15849–15854.

Büttner, D. & He, S.Y., 2009. Type III protein secretion in plant pathogenic bacteria. *Plant Physiology*, 150(4), pp.1656–64.

Carlson, C.S., Thomas, D.J., Eberle, M.A., Swanson, J.E., Livingston, R.J., Rieder, M.J. & Nickerson, D.A., 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome research*, 15(11), pp.1553–65.

Charlesworth, B., Charlesworth, D. & Barton, N.H., 2003. The Effects of Genetic and Geographic Structure on Neutral Variation. *Annual Review of Ecology, Evolution, and Systematics*, 34(1), pp.99–125.

Charlesworth, D., 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, 2(4), p.e64.

Chen, W., Provart, N. & Glazebrook, J., 2002. Expression profile matrix of Arabidopsis transcription factor genes suggests their putative functions in response to environmental stresses. *The Plant Cell Online*, 14(3), pp.559–574.

Choi, Y.-J., Shin, H.-D. & Thines, M., 2009. The host range of Albugo candida extends from Brassicaceae through Cleomaceae to Capparaceae. *Mycological Progress*, 8(4), pp.329–335.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M.F., Kellis, M., Lindblad-Toh, K. & Lander, E.S., 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49), pp.19428–33.

Cohen, S.N. & Chang, A.C.Y., 1973. Recircularization and Autonomous Replication of a Sheared R-Factor DNA Segment in Escherichia coli Transformants. *Proceedings of the National Academy of Sciences*, 70(5), pp.1293–1297.

Collins, S. & Meaux, J. De, 2009. Adaptation to different rates of environmental change in Chlamydomonas. *Evolution*, 63(11), pp.2952–65.

Cooper, A.J., Latunde-Dada, A.O., Woods-Tör, A., Lynn, J., Lucas, J.A., Crute, I.R. & Holub, E.B., 2008. Basic compatibility of Albugo candida in Arabidopsis thaliana and Brassica juncea causes broad-spectrum suppression of innate immunity. *Molecular Plant-Microbe Interactions*, 21(6), pp.745–56.

Cosson, P. & Sofer, L., 2010. A member of a new plant gene family encoding a meprin and TRAF homology (MATH) domain-containing protein is involved in restriction of long distance movement of plant viruses. *Plant Signaling & Behavior*, 5(10), pp.1321–3.

Costanza, R., Graumlich, L., Steffen, W., Crumley, C., Dearing, J., Hibbard, K., Leemans, R., Redman, C. & Schimel, D., 2007. Sustainability or Collapse: What Can We Learn from Integrating the History of Humans and the Rest of Nature? *AMBIO: A Journal of the Human Environment*, 36(7), pp.522–527.

Crick, F., 1958. On Protein Synthesis. *Symposia of the Society for Experimental Biology*.

Crow, J.F., 2010. Wright and Fisher on inbreeding and random drift. *Genetics*, 184(3), pp.609–11.

Dangl, J.L. & Jones, J.D., 2001. Plant pathogens and integrated defence responses to infection. *Nature*, 411(6839), pp.826–33.

Dangl, J.L. & McDowell, J.M., 2006. Two modes of pathogen recognition by plants. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23), pp.8575–6.

Dawkins, R. & Krebs, J.R., 1979. Arms Races between and within Species. *Proceedings of the Royal Society B: Biological Sciences*, 205(1161), pp.489–511.

DEFRA, 2009. UK Food Security Assessment : Our approach.

Deng, H.-W., Chen, W.-M. & Recker, R.R., 2001. Population Admixture: Detection by Hardy-Weinberg Test and Its Quantitative Effects on Linkage-Disequilibrium Methods for Localizing Genes Underlying Complex Traits. *Genetics*, 157(2), pp.885–897.

Devereux, S. & Edwards, J., 2004. Climate Change and Food Security. *IDS Bulletin*, 35(3), pp.22–30.

DeYoung, B.J., Bickle, K.L., Schrage, K.J., Muskett, P., Patel, K. & Clark, S.E., 2006. The CLAVATA1-related BAM1, BAM2 and BAM3 receptor kinase-like proteins are required for meristem function in Arabidopsis. *The Plant Journal*, 45(1), pp.1–16.

Dietz, T., Ostrom, E. & Stern, P., 2003. The struggle to govern the commons. *Science*, 302(5652), pp.1907–12.

Ding, J., Zhang, W., Jing, Z., Chen, J.-Q. & Tian, D., 2007. Unique pattern of R-gene variation within populations in Arabidopsis. *Molecular Genetics and Genomics*, 277(6), pp.619–629.

Dodds, P. & Rathjen, J., 2010. Plant immunity: towards an integrated view of plant–pathogen interactions. *Nature Reviews Genetics*, 11(8), pp.539–48.

Doerge, R., 2002. Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics*, 3(1), pp.43–52.

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H.O.K., Buffalo, V., Zerbino, D.R., Diekhans, M., Nguyen, N., Ariyaratne, P.N., Sung, W.-K., Ning, Z., Haimel, M., Simpson, J.T., Fonseca, N.A., Birol, İ., Docking, T.R., et al., 2011. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, 21(12), pp.2224–41.

Edgar, R., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1), pp.207–210.

Ehrenreich, I.M., Hanzawa, Y., Chou, L., Roe, J.L., Kover, P.X. & Purugganan, M.D., 2009. Candidate gene association mapping of Arabidopsis flowering time. *Genetics*, 183(1), pp.325–35.

Fahrig, L. & Merriam, G., 1994. Conservation of Fragmented Populations. *Conservation Biology*, 8(1), pp.50–59.

Falush, D., Stephens, M. & Pritchard, J.K., 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4), pp.1567–87.

Felix, G., Duran, J., Volko, S. & Boller, T., 1999. Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. *The Plant Journal*, 18(3), pp.265–76.

Felsenstein, J., 2002. PHYLIP (Phylogeny Inference Package) version 3.6a3. *Department of Genetics, University of Washington, Seattle*.

Feng, Y.-L., Lei, Y.-B., Wang, R.-F., Callaway, R.M., Valiente-Banuet, A., Inderjit, Li, Y.-P. & Zheng, Y.-L., 2009. Evolutionary tradeoffs for nitrogen allocation to photosynthesis versus cell walls in an invasive plant. *Proceedings of the National Academy of Sciences of the United States of America*, 106(6), pp.1853–6.

Fisher, R.A., 2012. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52(02), pp.399–433.

Flor, H.H., 1971. Current Status of the Gene-For-Gene Concept. *Annual Review of Phytopathology*, 9(1), pp.275–296.

Fradin, E.F., Abd-El-Haliem, A., Masini, L., van den Berg, G.C.M., Joosten, M.H.A.J. & Thomma, B.P.H.J., 2011. Interfamily transfer of tomato Ve1 mediates Verticillium resistance in Arabidopsis. *Plant Physiology*, 156(4), pp.2255–65.

François, O., Blum, M.G.B., Jakobsson, M. & Rosenberg, N. a, 2008. Demographic history of European populations of Arabidopsis thaliana. *PLoS Genetics*, 4(5), p.e1000075.

Fu, Y.X., 1994. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. *Genetics*, 138(4), pp.1375–86.

Gathorne-Hardy, F.. & Harcourt-Smith, W.E.., 2003. The super-eruption of Toba, did it cause a human bottleneck? *Journal of Human Evolution*, 45(3), pp.227–230.

Gavrilets, S., Acton, R. & Gravner, J., 2000. Dynamics of speciation and diversification in a metapopulation. *Evolution*, 54(5), pp.1493–1501.

Gene Ontology Consortium, 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(Database issue), pp.D258–61.

Georgia, J.C.A.P. of G.U. of, 2004. *The Hope, Hype, and Reality of Genetic Engineering : Remarkable Stories from Agriculture, Industry, Medicine, and the Environment: Remarkable Stories from Agriculture, Industry, Medicine, and the Environment (Google eBook)*, Oxford University Press.

Gilligan, C. a, 2008. Sustainable agriculture and plant diseases: an epidemiological perspective. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1492), pp.741–59.

Giraut, L., Falque, M., Drouaud, J., Pereira, L., Martin, O.C. & Mézard, C., 2011. Genome-wide crossover distribution in Arabidopsis thaliana meiosis reveals sex-specific patterns along chromosomes. *PLoS Genetics*, 7(11), p.e1002354.

Goda, H., Shimada, Y., Asami, T., Fujioka, S. & Yoshida, S., 2002. Microarray analysis of brassinosteroid-regulated genes in Arabidopsis. *Plant Physiology*, 130(3), pp.1319–34.

Goeddel, D. V., Kleid, D.G., Bolivar, F., Heyneker, H.L., Yansura, D.G., Crea, R., Hirose, T., Kraszewski, A., Itakura, K. & Riggs, A.D., 1979. Expression in Escherichia coli of chemically synthesized genes for human insulin. *Proceedings of the National Academy of Sciences*, 76(1), pp.106–110.

Gómez-Gómez, L. & Boller, T., 2000. FLS2: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. *Molecular Cell*, Volume 5(Issue 6), pp.1003–1011.

Gómez-Gómez, L., Felix, G. & Boller, T., 1999. A single locus determines sensitivity to bacterial flagellin in Arabidopsis thaliana. *The Plant Journal*, 18(December 1998), pp.277–284.

Govrin, E.M. & Levine, A., 2000. The hypersensitive response facilitates plant infection by the necrotrophic pathogen Botrytis cinerea. *Current Biology*, 10(13), pp.751–7.

Grant, M., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., Innes, R. & Dangl, J., 1995. Structure of the Arabidopsis RPM1 gene enabling dual specificity disease resistance. *Science*, 269(5225), pp.843–846.

Greilhuber, J., Borsch, T., Müller, K., Worberg, A., Porembski, S. & Barthlott, W., 2006. Smallest angiosperm genomes found in lentibulariaceae, with chromosomes of bacterial size. *Plant Biology*, 8(6), pp.770–7.

Griffiths, A., Gelbart, W., Miller, J. & Lewontin, R., 1999. *Recombination within a Gene*,

Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C. & Gelbart, W.M., 2000. *Accurate calculation of large map distances*, W. H. Freeman.

Griffiths-Jones, S., 2003. Rfam: an RNA family database. *Nucleic Acids Research*, 31(1), pp.439–441.

Gunz, P., Bookstein, F.L., Mittero-, P., Stadlmayr, A., Seidler, H., Gerhard, W., Weber, G.W., Kelly, D.J., Poolman, B., Gavin, H., Zeniou-meyer, M., Liu, Y., Olanich, M.E., Becherer, U., Bader, M. & Vitale, N., 2009. Early modern human diversity suggests subdivided population structure and a complex out-of-Africa scenario. *Proceedings of the National Academy of Sciences*, 106(20), pp.8398–8398.

Haefele, D.M. & Lindow, S.E., 1987. Flagellar Motility Confers Epiphytic Fitness Advantages upon Pseudomonas syringae. *Applied and Environmental Microbiology*, 53(10), pp.2528–33.

Hagenblad, J. & Nordborg, M., 2002. Sequence variation and haplotype structure surrounding the flowering time locus FRI in Arabidopsis thaliana. *Genetics*, 161(1), pp.289–98.

Hall, M.C. & Willis, J.H., 2006. Divergent selection on flowering time contributes to local adaptation in Mimulus guttatus populations. *Evolution*, 60(12), pp.2466–2477.

Hanfstingl, U., Berry, A., Kellogg, E.A., Costa, J.T., Rüdiger, W. & Ausubel, F.M., 1994. Haplotypic divergence coupled with lack of diversity at the Arabidopsis thaliana alcohol dehydrogenase locus: roles for both balancing and directional selection? *Genetics*, 138(3), pp.811–28.

Hanski, I., 1998. Metapopulation dynamics. *Nature*, 396(6706), pp.41–49.

Hardy, G.H., 2003. Mendelian proportions in a mixed population. 1908. *The Yale journal of biology and medicine*, 76(2), pp.79–80.

Hardy, O.J. & Vekemans, X., 1999. Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity*, 83 (Pt 2)(April), pp.145–54.

Hasan, H., 2004. *Mendel and the Laws of Genetics*, The Rosen Publishing Group.

Heil, M., 2002. Fitness costs of induced resistance: emerging experimental support for a slippery concept. *Trends in Plant Science*, 7(2), pp.61–67.

Heil, M., Hilpert, A., Kaiser, W. & Linsenmair, K.E., 2000. Reduced growth and seed set following chemical induction of pathogen defence: does systemic acquired resistance (SAR) incur allocation costs? *Journal of Ecology*, 88(4), pp.645–654.

Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., Sella, G. & Przeworski, M., 2011. Classic selective sweeps were rare in recent human evolution. *Science*, 331(6019), pp.920–4.

Hewitt, G., 1999. Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, 68(1-2), pp.87–112.

Hoegh-Guldberg, O., Mumby, P.J., Hooten, A.J., Steneck, R.S., Greenfield, P., Gomez, E., Harvell, C.D., Sale, P.F., Edwards, A.J., Caldeira, K., Knowlton, N., Eakin, C.M., Iglesias-Prieto, R., Muthiga, N., Bradbury, R.H., Dubi, A. & Hatziolos, M.E., 2007. Coral reefs under rapid climate change and ocean acidification. *Science*, 318(5857), pp.1737–42.

Hoffmann, M., 2002. Biogeography of Arabidopsis thaliana (L.) Heynh (Brassicacaeae). *Journal of Biogeography*, pp.125–134.

Holderegger, R., Kamm, U. & Gugerli, F., 2006. Adaptive vs. neutral genetic diversity: implications for landscape genetics. *Landscape Ecology*, 21(6), pp.797–807.

Holub, E., 2001. The arms race is ancient history in Arabidopsis, the wildflower. *Nature Reviews Genetics*, 2(July), pp.516–527.

Holub, E.B., 2008. Natural history of Arabidopsis thaliana and oomycete symbioses. *European Journal of Plant Pathology*, 122(1), pp.91–109.

Hoon, M. de, Imoto, S., Nolan, J. & Miyano, S., 2004. Open source clustering software. *Bioinformatics*.

Hord, C.L.H., Chen, C., Deyoung, B.J., Clark, S.E. & Ma, H., 2006. The BAM1/BAM2 receptor-like kinases are important regulators of Arabidopsis early anther development. *The Plant Cell*, 18(7), pp.1667–80.

Horton, M.W., Hancock, A.M., Huang, Y.S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N.W., Platt, A., Sperone, F.G., Vilhjálmsson, B.J., Nordborg, M., Borevitz, J.O. & Bergelson, J.,

2012. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics*, 44(2), pp.212–6.

Innan, H. & Stephan, W., 2000. The Coalescent in an Exponentially Growing Metapopulation and Its Application to Arabidopsis thaliana. *Genetics*, 155(4), pp.2015–2019.

Innan, H., Terauchi, R. & Miyashita, N.T., 1997. Microsatellite polymorphism in natural populations of the wild plant Arabidopsis thaliana. *Genetics*, 146(4), pp.1441–52.

Jackson, J.B., Kirby, M.X., Berger, W.H., Bjorndal, K.A., Botsford, L.W., Bourque, B.J., Bradbury, R.H., Cooke, R., Erlandson, J., Estes, J.A., Hughes, T.P., Kidwell, S., Lange, C.B., Lenihan, H.S., Pandolfi, J.M., Peterson, C.H., Steneck, R.S., Tegner, M.J. & Warner, R.R., 2001. Historical overfishing and the recent collapse of coastal ecosystems. *Science*, 293(5530), pp.629–37.

Johnson, J.M., Edwards, S., Shoemaker, D. & Schadt, E.E., 2005. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends in Genetics*, 21(2), pp.93–102.

Jones, J.D.G. & Dangl, J.L., 2006. The plant immune system. *Nature*, 444(7117), pp.323–9.

Kanz, C., Aldebert, P. & Althorpe, N., 2005. The EMBL nucleotide sequence database. *Nucleic Acids Research*, 32(Database issue), pp.D27–30.

Karp, R.M., 2003. Large scale reconstruction of haplotypes from genotype data. In *Proceedings of the seventh annual international conference on Computational Molecular Biology - RECOMB '03*. New York, New York, USA: ACM Press, pp. 104–113.

Kawecki, T.J. & Ebert, D., 2004. Conceptual issues in local adaptation. *Ecology Letters*, 7(12), pp.1225–1241.

Keurentjes, J.J.B., Bentsink, L., Alonso-Blanco, C., Hanhart, C.J., Blankestijn-De Vries, H., Effgen, S., Vreugdenhil, D. & Koornneef, M., 2007. Development of a near-isogenic line population of Arabidopsis thaliana and comparison of mapping power with a recombinant inbred line population. *Genetics*, 175(2), pp.891–905.

Kim, M.G., da Cunha, L., McFall, A.J., Belkhadir, Y., DebRoy, S., Dangl, J.L. & Mackey, D., 2005. Two Pseudomonas syringae type III effectors inhibit RIN4-regulated basal defense in Arabidopsis. *Cell*, 121(5), pp.749–59.

Kim, S., Plagnol, V., Hu, T.T., Toomajian, C., Clark, R.M., Ossowski, S., Ecker, J.R., Weigel, D. & Nordborg, M., 2007. Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics*, 39(9), pp.1151–5.

Kim, S.H., Kwon, S. Il, Saha, D., Anyanwu, N.C. & Gassmann, W., 2009. Resistance to the Pseudomonas syringae effector HopA1 is governed by the TIR-NBS-LRR protein RPS6 and is enhanced by mutations in SRFR1. *Plant Physiology*, 150(4), pp.1723–32.

Kim, T.-H., Kunz, H.-H., Bhattacharjee, S., Hauser, F., Park, J., Engineer, C., Liu, A., Ha, T., Parker, J.E., Gassmann, W. & Schroeder, J.I., 2012. Natural variation in small molecule-induced TIR-NB-LRR signaling induces root growth arrest via EDS1- and PAD4-complexed R protein VICTR in Arabidopsis. *The Plant Cell*, 24(12), pp.5177–92.

Kimura, M. & Ota, T., 1973. The age of a neutral mutant persisting in a finite population. *Genetics*, 75(1), pp.199–212.

Kimura, M. & Weiss, G., 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, (480), pp.561–576.

King, M. & Wilson, a., 1975. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184), pp.107–116.

Kniskern, J.M., Traw, M.B. & Bergelson, J., 2007. Salicylic acid and jasmonic acid signaling defense pathways reduce natural bacterial diversity on Arabidopsis thaliana. *Molecular Plant-Microbe Interactions*, 20(12), pp.1512–22.

Kobayashi, D.Y., Tamaki, S.J. & Keen, N.T., 1989. Cloned avirulence genes from the tomato pathogen Pseudomonas syringae pv. tomato confer cultivar specificity on soybean. *Proceedings of the National Academy of Sciences of the United States of America*, 86(1), pp.157–61.

Korte, A. & Farlow, A., 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, 9(1), p.29.

Kover, P.X. & Schaal, B.A., 2002. Genetic variation for disease resistance and tolerance among Arabidopsis thaliana accessions. *Proceedings of the National Academy of Sciences of the United States of America*, 99(17), pp.11270–4.

Kover, P.X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I.M., Purugganan, M.D., Durrant, C. & Mott, R., 2009. A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genetics*, 5(7), p.e1000551.

Kreitman, M., 2000. Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics*, 1, pp.539–59.

Kreps, J.A., Wu, Y., Chang, H.-S., Zhu, T., Wang, X. & Harper, J.F., 2002. Transcriptome changes for Arabidopsis in response to salt, osmotic, and cold stress. *Plant Physiology*, 130(4), pp.2129–41.

Kroymann, J., Donnerhacke, S., Schnabelrauch, D. & Mitchell-Olds, T., 2003. Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. *Proceedings of the National Academy of Sciences of the United States of America*, 100 Suppl , pp.14587–92.

Kruglyak, L., 2008. The road to genome-wide association studies. *Nature Reviews Genetics*, 9(4), pp.314–8.

Lander, E.S., 2011. Initial impact of the sequencing of the human genome. *Nature*, 470(7333), pp.187–97.

Lander, E.S. & Green, P., 1987. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, 84(8), pp.2363–2367.

Landschulz, W., Johnson, P. & McKnight, S., 1988. The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science*, 240(4860), pp.1759–1764.

Lawson, D.J., Hellenthal, G., Myers, S. & Falush, D., 2012. Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1), p.e1002453.

Levin, D.A., 1995. Metapopulations: an arena for local speciation. *Journal of Evolutionary Biology*, 8(5), pp.635–644.

Lewis, J.D., Wu, R., Guttman, D.S. & Desveaux, D., 2010. Allele-specific virulence attenuation of the Pseudomonas syringae HopZ1a type III effector via the Arabidopsis ZAR1 resistance protein. *PLoS Genetics*, 6(4), p.e1000894.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, Shengting, Shan, G., Kristiansen, K., Li, Songgang, Yang, H., Wang, Jian & Wang, Jun, 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2), pp.265–72.

Li, Y., Zheng, L., Corke, F., Smith, C. & Bevan, M.W., 2008. Control of final seed and organ size by the DA1 gene family in Arabidopsis thaliana. *Genes & Development*, 22(10), pp.1331–6.

Little, D., Gouhier-Darimont, C., Bruessow, F. & Reymond, P., 2007. Oviposition by pierid butterflies triggers defense responses in Arabidopsis. *Plant Physiology*, 143(2), pp.784–800.

Liu, B.H., 1997. *Statistical Genomics: Linkage, Mapping, and QTL Analysis*, CRC Press.

Loake, G. & Grant, M., 2007. Salicylic acid in plant defence--the players and protagonists. *Current Opinion in Plant Biology*, 10(5), pp.466–72.

Lynch, M. & Conery, J.S., 2003. The origins of genome complexity. *Science*, 302(5649), pp.1401–4.

Mackay, T., Stone, E. & Ayroles, J., 2009. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*.

Macnair, M., 1993. The genetics of metal tolerance in vascular plants. *The New Phytologist*, (49), pp.541–559.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. The Regents of the University of California.

Manolio, T.A., Brooks, L.D. & Collins, F.S., 2008. A HapMap harvest of insights into the genetics of common disease. *The Journal of Clinical Investigation*, 118(5), pp.1590–605.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., et al., 2009. Finding the missing heritability of complex diseases. *Nature*, 461(7265), pp.747–53.

Marsh, D.M. & Trenham, P.C., 2001. Metapopulation Dynamics and Amphibian Conservation. *Conservation Biology*, 15(1), pp.40–49.

Masel, J., 2011. Genetic drift. *Current Biology*, 21(20), pp.R837–8.

Maston, G.A., Evans, S.K. & Green, M.R., 2006. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*, 7, pp.29–59.

Mauricio, R., Stahl, E. a, Korves, T., Tian, D., Kreitman, M. & Bergelson, J., 2003. Natural selection for polymorphism in the disease resistance gene Rps2 of Arabidopsis thaliana. *Genetics*, 163(2), pp.735–46.

McDowell, J.M., 1998. Intragenic Recombination and Diversifying Selection Contribute to the Evolution of Downy Mildew Resistance at the RPP8 Locus of Arabidopsis. *The Plant Cell Online*, 10(11), pp.1861–1874.

McEntyre, J., Ostell, J. & Mizrachi, I., 2007. *GenBank: The Nucleotide Sequence Database*, National Center for Biotechnology Information (US).

De Meaux, J. & Mitchell-Olds, T., 2003. Evolution of plant resistance at the molecular level: ecological context of species interactions. *Heredity*, 91(4), pp.345–52.

Meyerowitz, E.M. & Pruitt, R.E., 1985. Arabidopsis thaliana and Plant Molecular Genetics. *Science*, 229(4719), pp.1214–8.

Michaels, S.D., He, Y., Scortecci, K.C. & Amasino, R.M., 2003. Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 100(17), pp.10102–7.

Mitchell-Olds, T., 2001. Arabidopsis thaliana and its wild relatives: a model system for ecology and evolution. *Trends in Ecology & Evolution*, 16(12), pp.693–700.

Mitchell-Olds, T., Feder, M. & Wray, G., 2008. Evolutionary and ecological functional genomics. *Heredity*, 100(2), pp.101–2.

Mitchell-Olds, T., Willis, J. & Goldstein, D., 2007. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics*, 8(11), pp.845–56.

Monaghan, J. & Zipfel, C., 2012. Plant pattern recognition receptor complexes at the plasma membrane. *Current Opinion in Plant Biology*, 15(4), pp.349–57.

Montillet, J.-L., Leonhardt, N., Mondy, S., Tranchimand, S., Rumeau, D., Boudsocq, M., Garcia, A.V., Douki, T., Bigeard, J., Laurière, C., Chevalier, A., Castresana, C. & Hirt, H., 2013. An abscisic acid-independent oxylipin pathway controls stomatal closure and immune defense in Arabidopsis. *PLoS Biology*, 11(3), p.e1001513.

Morgan, T.H., 1911. Random segregation versus coupling in Mendelian inheritance. *Science*, 34(873), p.384.

Morgan, T.H., 1921. The physical basis of heredity. *Zeitschrift für Induktive Abstammungs- und Vererbungslehre*, 26(1-2), pp.176–178.

Mosher, S., Seybold, H., Rodriguez, P., Stahl, M., Davies, K. a, Dayaratne, S., Morillo, S. a, Wierzba, M., Favery, B., Keller, H., Tax, F.E. & Kemmerling, B., 2013. The tyrosine-sulfated peptide receptors PSKR1 and PSY1R modify the immunity of Arabidopsis to biotrophic and necrotrophic pathogens in an antagonistic manner. *The Plant Journal*, 73(3), pp.469–82.

Munkres, J., 1957. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial & Applied Mathematics*, 5(1), pp.32–38.

Mur, L. a J., Kenton, P., Lloyd, A.J., Ougham, H. & Prats, E., 2008. The hypersensitive response; the centenary is upon us but how much do we know? *Journal of Experimental Botany*, 59(3), pp.501–20.

Nemri, A., Atwell, S., Tarone, A.M., Huang, Y.S., Zhao, K., Studholme, D.J., Nordborg, M. & Jones, J.D.G., 2010. Genome-wide survey of Arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 107(22), pp.10302–7.

Nimchuk, Z., Eulgem, T., Holt, B.F. & Dangl, J.L., 2003. Recognition and response in the plant immune system. *Annual Review of Genetics*, 37, pp.579–609.

Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J.N., Noyes, T., Oefner, P.J., Stahl, E. a & Weigel, D., 2002. The extent of linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics*, 30(2), pp.190–3.

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N. a, Shah, C., Wall, J.D., Wang, J., et al., 2005. The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biology*, 3(7), p.e196.

Ohta, T., 1992. The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*.

Olivier, M., Aggarwal, A., Allen, J., Almendras, A.A., Bajorek, E.S., Beasley, E.M., Brady, S.D., Bushard, J.M., Bustos, V.I., Chu, A., Chung, T.R., Witte, A. De, Denys, M.E., Dominguez, R., Fang, N.Y., Foster, B.D., Freudenberg, R.W., Hadley, D., Hamilton, L.R., et al., 2001. A High-Resolution Radiation Hybrid Map of the Human Genome Draft Sequence. *Science*, 291(February), pp.1298–1302.

Pandolfi, J.M., Bradbury, R.H., Sala, E., Hughes, T.P., Bjorndal, K.A., Cooke, R.G., McArdle, D., McClenachan, L., Newman, M.J.H., Paredes, G., Warner, R.R. & Jackson, J.B.C., 2003. Global trajectories of the long-term decline of coral reef ecosystems. *Science*, 301(5635), pp.955–8.

Panopoulos, N.J., 1974. Role of Flagellar Motility in the Invasion of Bean Leaves by Pseudomonas phaseolicola. *Phytopathology*, 64(11), p.1389.

Parker, I.M. & Gilbert, G.S., 2004. the Evolutionary Ecology of Novel Plant-Pathogen Interactions. *Annual Review of Ecology, Evolution, and Systematics*, 35(1), pp.675–700.

Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S. & Reich, D., 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097), pp.1103–8.

Pearson, T.A. & Manolio, T.A., 2008. How to interpret a genome-wide association study. *The Journal of the American Medical Association*, 299(11), pp.1335–44.

Pigliucci, M., 1998. Ecological and evolutionary genetics of Arabidopsis. *Trends in Plant Science*, 13, pp.485–489.

Platt, A., Horton, M., Huang, Y.S., Li, Y., Anastasio, A.E., Mulyati, N.W., Agren, J., Bossdorf, O., Byers, D., Donohue, K., Dunning, M., Holub, E.B., Hudson, A., Le Corre, V., Loudet, O., Roux, F., Warthmann, N., Weigel, D., Rivero, L., et al., 2010. The scale of population structure in Arabidopsis thaliana. *PLoS Genetics*, 6(2), p.e1000843.

Ploch, S., Choi, Y.-J., Rost, C., Shin, H.-D., Schilling, E. & Thines, M., 2010. Evolution of diversity in Albugo is driven by high host specificity and multiple speciation events on closely related Brassicaceae. *Molecular Phylogenetics and Evolution*, 57(2), pp.812–20.

Plomin, R., Haworth, C.M.A. & Davis, O.S.P., 2009. Common disorders are quantitative traits. *Nature Reviews Genetics*, 10(12), pp.872–8.

Popescu, S.C., Popescu, G. V, Bachan, S., Zhang, Z., Gerstein, M., Snyder, M. & Dinesh-Kumar, S.P., 2009. MAPK target networks in Arabidopsis thaliana revealed using functional protein microarrays. *Genes & Development*, 23(1), pp.80–92.

Pritchard, J.K., Pickrell, J.K. & Coop, G., 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, 20(4), pp.R208–15.

Pritchard, J.K., Stephens, M. & Donnelly, P., 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155(2), pp.945–59.

Provart, N. & Zhu, T., 2003. A browser-based functional classification SuperViewer for Arabidopsis genomics. *Currents in Computational Molecular Biology*, pp.3–4.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. & Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13(1), p.341.

Ramos-Onsins, S.E., 2004. Multilocus Analysis of Variation and Speciation in the Closely Related Species Arabidopsis halleri and A. lyrata. *Genetics*, 166(1), pp.373–388.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. & Lander, E.S., 2001. Linkage disequilibrium in the human genome. *Nature*, 411(6834), pp.199–204.

Replogle, A., Wang, J., Paolillo, V., Smeda, J., Kinoshita, A., Durbak, A., Tax, F.E., Wang, X., Sawa, S. & Mitchum, M.G., 2013. Synergistic interaction of CLAVATA1, CLAVATA2, and RECEPTOR-LIKE PROTEIN KINASE 2 in cyst nematode parasitism of Arabidopsis. *Molecular Plant-Microbe Interactions*, 26(1), pp.87–96.

Robert, V., Tchuinkam, T., Mulder, B., Bodo, J.M., Verhave, J.P., Carnevale, P. & Nagel, R.L., 1996. Effect of the sickle cell trait status of gametocyte carriers of Plasmodium falciparum on infectivity to anophelines. *The American Journal of Tropical Medicine and Hygiene*, 54(2), pp.111–3.

Robusto, C., 1957. The cosine-haversine formula. *American Mathematical Monthly*, Vol. 64(No. 1 (Jan., 1957)), pp.pp. 38–40.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P., 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, 242(1), pp.84–9.

Rose, L., Atwell, S., Grant, M. & Holub, E.B., 2012. Parallel Loss-of-Function at the RPM1 Bacterial Resistance Locus in Arabidopsis thaliana. *Frontiers in Plant Science*, 3(December), p.287.

Rose, L.E., 2004. The Maintenance of Extreme Amino Acid Diversity at the Disease Resistance Gene, RPP13, in Arabidopsis thaliana. *Genetics*, 166(3), pp.1517–1527.

Rose, M. & Charlesworth, B., 1980. A test of evolutionary theories of senescence. *Nature*, 287(5778), pp.141–142.

Rosegrant, M.W. & Cline, S.A., 2003. Global food security: challenges and policies. *Science*, 302(5652), pp.1917–9.

Roux, F., Giancola, S., Durand, S. & Reboud, X., 2006. Building of an experimental cline with Arabidopsis thaliana to estimate herbicide fitness cost. *Genetics*, 173(2), pp.1023–31.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J. V, Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. & Lander, E.S., 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909), pp.832–7.

Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D. & Lander, E.S., 2006. Positive natural selection in the human lineage. *Science*, 312(5780), pp.1614–20.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S. a, Gaudet, R., Schaffner, S.F., Lander, E.S., Frazer, K. a, Ballinger, D.G., Cox, D.R., Hinds, D. a, Stuve, L.L., Gibbs, R. a, Belmont, J.W., et al., 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164), pp.913–8.

Sakamoto, T., Deguchi, M., Brustolini, O.J.B., Santos, A.A., Silva, F.F. & Fontes, E.P.B., 2012. The tomato RLK superfamily: phylogeny and functional predictions about the role of the LRRII-RLK subfamily in antiviral defense. *BMC Plant Biology*, 12, p.229.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M. & Smith, M., 1977. Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature*, 265(5596), pp.687–695.

Schena, M., Shalon, D., Davis, R.W. & Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235), pp.467–70.

Scholthof, K.-B.G., 2007. The disease triangle: pathogens, the environment and society. *Nature reviews. Microbiology*, 5(2), pp.152–6.

Schwartz, J., 2009. *In Pursuit of the Gene: From Darwin to DNA (Google eBook)*, Harvard University Press.

Scourse, J.D. & Austin, R.M., 1995. Palaeotidal modelling of continental shelves: marine implications of a land-bridge in the Strait of Dover during the Holocene and Middle Pleistocene. *Geological Society, London, Special Publications*, 96(1), pp.75–88.

SHAPIRO, S.S. & WILK, M.B., 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), pp.591–611.

Sharbel, T.F., Haubold, B. & Mitchell-Olds, T., 2000. Genetic isolation by distance in Arabidopsis thaliana: biogeography and postglacial colonization of Europe. *Molecular Ecology*, 9(12), pp.2109–18.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J.G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J.F. & Campbell, H., 2011. Abundant pleiotropy in human complex diseases and traits. *American Journal of Human Genetics*, 89(5), pp.607–18.

Smith, J.M. & Haigh, J., 1974. The hitch-hiking effect of a favourable gene. *Genetical Research*, 23(01), p.23.

Snape, J. & Lawrence, M., 1971. The breeding system of Arabidopsis thaliana. *Heredity*, pp.299–302.

Sokal, R., 1958. *A statistical method for evaluating systematic relationships*,

Stahl, E. a, Dwyer, G., Mauricio, R., Kreitman, M. & Bergelson, J., 1999. Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. *Nature*, 400(6745), pp.667–71.

Stahl, Y., Grabowski, S., Bleckmann, A., Kühnemuth, R., Weidtkamp-Peters, S., Pinto, K.G., Kirschner, G.K., Schmid, J.B., Wink, R.H., Hülsewede, A., Felekyan, S., Seidel, C.A.M. & Simon, R., 2013. Moderation of Arabidopsis root stemness by CLAVATA1 and ARABIDOPSIS CRINKLY4 receptor kinase complexes. *Current Biology*, 23(5), pp.362–71.

Stein, L.D., 2010. The case for cloud computing in genome informatics. *Genome Biology*, 11(5), p.207.

Sujansky, W., 2001. Heterogeneous database integration in biomedicine. *Journal of Biomedical Informatics*, 34(4), pp.285–98.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), pp.585–95.

The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), pp.1061–73.

The International HapMap Consortium, 2005. A haplotype map of the human genome. *Nature*, 437(7063), pp.1299–320.

The International HapMap Consortium, 2003. The International HapMap Project. *Nature*, 426(6968), pp.789–96.

The Wellcome Trust Case Control Consortium, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), pp.661–78.

Thrall, P.H. & Burdon, J.J., 2003. Evolution of virulence in a plant host-pathogen metapopulation. *Science*, 299(5613), pp.1735–7.

Tian, D., Traw, M.B., Chen, J.Q., Kreitman, M. & Bergelson, J., 2003. Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana. *Nature*, 423(6935), pp.74–7.

Todesco, M., Balasubramanian, S., Hu, T.T., Traw, M.B., Horton, M., Epple, P., Kuhns, C., Sureshkumar, S., Schwartz, C., Lanz, C., Laitinen, R. a E., Huang, Y., Chory, J., Lipka, V., Borevitz, J.O., Dangl, J.L., Bergelson, J., Nordborg, M. & Weigel, D., 2010. Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana. *Nature*, 465(7298), pp.632–6.

Tsuchihashi, Z. & Dracopoli, N.C., 2002. Progress in high throughput SNP genotyping methods. *The Pharmacogenomics Journal*, 2(2), pp.103–110.

Vitousek, P.M., 1997. Human Domination of Earth's Ecosystems. *Science*, 277(5325), pp.494–499.

Wackernagel, M. & Rees, W.E., 2013. *Our Ecological Footprint: Reducing Human Impact on the Earth*, New Society Publishers.

Wall, J.D. & Pritchard, J.K., 2003. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 4(8), pp.587–97.

Wang, Y., Zhang, Y., Wang, Z., Zhang, X. & Yang, S., 2013. A missense mutation in CHS1, a TIR-NB protein, induces chilling sensitivity in Arabidopsis. *The Plant Journal*, 75(4), pp.553–65.

Weigel, D. & Mott, R., 2009. The 1001 genomes project for Arabidopsis thaliana. *Genome Biology*, 10(5), p.107.

Weinberg, W., 1908. The demonstration of heredity in man. *Boyer SH, trans (1963) Papers on human genetics*. ….

Weir, B. & Cockerham, C., 1984. Estimating F-statistics for the analysis of population structure. *Evolution*, Vol. 38(No. 6 (Nov., 1984)), pp.pp. 1358–1370.

Wender, N.J., Polisetty, C.R. & Donohue, K., 2005. Density-dependent processes influencing the evolutionary dynamics of dispersal: a functional analysis of seed dispersal in Arabidopsis thaliana (Brassicaceae). *American Journal of Botany*, 92(6), pp.960–71.

Wigginton, J.E., Cutler, D.J. & Abecasis, G.R., 2005. A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics*, 76(5), pp.887–93.

Williams, T.N., Mwangi, T.W., Wambua, S., Alexander, N.D., Kortok, M., Snow, R.W. & Marsh, K., 2005. Sickle cell trait and the risk of Plasmodium falciparum malaria and other childhood diseases. *The Journal of Infectious Diseases*, 192(1), pp.178–86.

Wright, S., 1950. Genetical structure of populations. *Nature*, 166(4215), pp.247–9.

Wright, S., 1938. Size of population and breeding structure in relation to evolution. *Science*, 87, pp.430–431.

Yelin, R., Dahary, D., Sorek, R., Levanon, E.Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., Nemzer, S., Pinner, E., Walach, S., Bernstein, J., Savitsky, K. & Rotman, G., 2003. Widespread occurrence of antisense transcription in the human genome. *Nature Biotechnology*, 21(4), pp.379–86.

Zahiri, J., Mahdevar, G., Nowzari-Dalini, A., Ahrabian, H. & Sadeghi, M., 2010. A novel efficient dynamic programming algorithm for haplotype block partitioning. *Journal of Theoretical Biology*, 267(2), pp.164–70.

Zhao, K., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P. & Nordborg, M., 2007. An Arabidopsis example of association mapping in structured samples. *PLoS Genetics*, 3(1), p.e4.

Zipfel, C., Kunze, G., Chinchilla, D., Caniard, A., Jones, J.D.G., Boller, T. & Felix, G., 2006. Perception of the bacterial PAMP EF-Tu by the receptor EFR restricts Agrobacterium-mediated transformation. *Cell*, 125(4), pp.749–60.

APPENDIX 1: UK-WIDE LONG-DISTANCE GENOTYPIC SIMILARITY

A1.1 P<=0.05

Accession 1	Lat.	Lon.	Accession 2	Lat.	Lon.	p-value
02B6	55.9218	-3.17108	Cnt_1	51.3	1.1	0.017
02B6	55.9218	-3.17108	Crl_1	54.9	-2.9	0.039
02B6	55.9218	-3.17108	EM_183	51.3	0.5	0.017
02B6	55.9218	-3.17108	Ema_1A	51.3	0.5	0.003
02B6	55.9218	-3.17108	HR_10	51.4083	-0.6383	0.010
02B6	55.9218	-3.17108	PHW_14	51.2878	0.0565	0.010
02B6	55.9218	-3.17108	UKID103	51.8	-0.5	0.001
02B6	55.9218	-3.17108	UKID17	51.4	0.1	0.003
02B6	55.9218	-3.17108	UKID28	52.3	-1.7	0.025
02B6	55.9218	-3.17108	UKID87	50.8	-0.7	0.010
02B6	55.9218	-3.17108	UKNW6_078	54.4	-3	0.036
02B6	55.9218	-3.17108	UKNW6_210	54.4	-3	0.036
02B6	55.9218	-3.17108	UKNW6_410	54.7	-3.4	0.047
02B6	55.9218	-3.17108	UKNW9_025	54.6	-3.1	0.019
02B6	55.9218	-3.17108	UKSE6_032	51.3	0.5	0.017
02B6	55.9218	-3.17108	UKSE6_350	51.3	0.4	0.017
02B6	55.9218	-3.17108	UKSE6_544	51.3	1.1	0.017
02B6	55.9218	-3.17108	UKSE6_618	51.1	0.4	0.010
02B6	55.9218	-3.17108	UKSW6_025	50.4	-4.7	0.020
02B6	55.9218	-3.17108	UKSW6_070	50.4	-4.7	0.020
09A3	55.9212	-3.17857	UKID39	55.4	-2.8	0.042
12A1	55.8877	-3.16377	13B5	55.8858	-3.16015	0.047
12A1	55.8877	-3.16377	Cnt_1	51.3	1.1	0.010
12A1	55.8877	-3.16377	Crl_1	54.9	-2.9	0.039
12A1	55.8877	-3.16377	Ema_1A	51.3	0.5	0.010
12A1	55.8877	-3.16377	HR_10	51.4083	-0.6383	0.017
12A1	55.8877	-3.16377	PHW_14	51.2878	0.0565	0.017
12A1	55.8877	-3.16377	UKID103	51.8	-0.5	0.003
12A1	55.8877	-3.16377	UKID17	51.4	0.1	0.010
12A1	55.8877	-3.16377	UKID87	50.8	-0.7	0.003
12A1	55.8877	-3.16377	UKNW6_078	54.4	-3	0.036
12A1	55.8877	-3.16377	UKNW6_210	54.4	-3	0.036
12A1	55.8877	-3.16377	UKNW9_025	54.6	-3.1	0.019
12A1	55.8877	-3.16377	UKSE6_350	51.3	0.4	0.001
12A1	55.8877	-3.16377	UKSE6_618	51.1	0.4	0.017
12A1	55.8877	-3.16377	UKSW6_025	50.4	-4.7	0.020
12A1	55.8877	-3.16377	UKSW6_070	50.4	-4.7	0.020
13B5	55.8858	-3.16015	Crl_1	54.9	-2.9	0.039
13B5	55.8858	-3.16015	NFA_8	51.4083	-0.6383	0.017
13B5	55.8858	-3.16015	Sq_1	51.4083	-0.6383	0.010
13B5	55.8858	-3.16015	UKID103	51.8	-0.5	0.024
13B5	55.8858	-3.16015	UKNW6_078	54.4	-3	0.036
13B5	55.8858	-3.16015	UKNW6_410	54.7	-3.4	0.047
13B5	55.8858	-3.16015	UKNW9_025	54.6	-3.1	0.010
13B5	55.8858	-3.16015	UKSW6_025	50.4	-4.7	0.020
13B5	55.8858	-3.16015	UKSW6_070	50.4	-4.7	0.020
Alst_1	54.8	-2.4333	Edburgh_8	55.9681	-3.21833	0.047
Alst_1	54.8	-2.4333	HR_10	51.4083	-0.6383	0.029
Alst_1	54.8	-2.4333	NFA_8	51.4083	-0.6383	0.010

Accession 1	Lat.	Lon.	Accession 2	Lat.	Lon.	p-value
Alst_1	54.8	-2.4333	Sis_1	51.1	0.6	0.024
Alst_1	54.8	-2.4333	Sq_1	51.4083	-0.6383	0.029
Alst_1	54.8	-2.4333	UKNW6_078	54.4	-3	0.017
Alst_1	54.8	-2.4333	UKNW6_410	54.7	-3.4	0.042
Alst_1	54.8	-2.4333	UKNW9_025	54.6	-3.1	0.022
Alst_1	54.8	-2.4333	UKSE6_618	51.1	0.4	0.024
Alst_1	54.8	-2.4333	UKSW6_025	50.4	-4.7	0.017
Alst_1	54.8	-2.4333	UKSW6_070	50.4	-4.7	0.017
Alst_1	54.8	-2.4333	Ullapool4	57.9	-5.1525	0.029
Boot_1	54.4	-3.2667	Poo_1	54.6	-2.8	0.049
Boot_1	54.4	-3.2667	UKNW6_425	54.7	-3.4	0.016
Boot_1	54.4	-3.2667	UKNW9_010	54.6	-3.1	0.016
Cnt_1	51.3	1.1	Crl_1	54.9	-2.9	0.010
Cnt_1	51.3	1.1	Ema_1A	51.3	0.5	0.026
Cnt_1	51.3	1.1	Hil_1	51	-1.5	0.011
Cnt_1	51.3	1.1	HR_10	51.4083	-0.6383	0.012
Cnt_1	51.3	1.1	HR_5	51.4083	-0.6383	0.047
Cnt_1	51.3	1.1	NFA_10	51.4083	-0.6383	0.047
Cnt_1	51.3	1.1	NFA_8	51.4083	-0.6383	0.012
Cnt_1	51.3	1.1	Sis_1	51.1	0.6	0.046
Cnt_1	51.3	1.1	Sq_1	51.4083	-0.6383	0.047
Cnt_1	51.3	1.1	UKID103	51.8	-0.5	0.012
Cnt_1	51.3	1.1	UKID17	51.4	0.1	0.024
Cnt_1	51.3	1.1	UKID35	51.3	0.9	0.048
Cnt_1	51.3	1.1	UKID55	53	-1.1	0.030
Cnt_1	51.3	1.1	UKID87	50.8	-0.7	0.003
Cnt_1	51.3	1.1	UKNW6_078	54.4	-3	0.012
Cnt_1	51.3	1.1	UKNW6_210	54.4	-3	0.004
Cnt_1	51.3	1.1	UKNW6_410	54.7	-3.4	0.024
Cnt_1	51.3	1.1	UKNW9_025	54.6	-3.1	0.024
Cnt_1	51.3	1.1	UKSE6_350	51.3	0.4	0.022
Cnt_1	51.3	1.1	UKSE6_556	51.3	1.1	0.047
Cnt_1	51.3	1.1	UKSE6_618	51.1	0.4	0.023
Cnt_1	51.3	1.1	UKSW6_025	50.4	-4.7	0.012
Cnt_1	51.3	1.1	UKSW6_070	50.4	-4.7	0.025
Crl_1	54.9	-2.9	Edburgh_8	55.9681	-3.21833	0.047
Crl_1	54.9	-2.9	Ema_1A	51.3	0.5	0.024
Crl_1	54.9	-2.9	Hil_1	51	-1.5	0.025
Crl_1	54.9	-2.9	HR_10	51.4083	-0.6383	0.012
Crl_1	54.9	-2.9	NFA_8	51.4083	-0.6383	0.002
Crl_1	54.9	-2.9	Sis_1	51.1	0.6	0.010
Crl_1	54.9	-2.9	Sq_1	51.4083	-0.6383	0.025
Crl_1	54.9	-2.9	UKID103	51.8	-0.5	0.029
Crl_1	54.9	-2.9	UKID17	51.4	0.1	0.025
Crl_1	54.9	-2.9	UKID35	51.3	0.9	0.024
Crl_1	54.9	-2.9	UKID55	53	-1.1	0.007
Crl_1	54.9	-2.9	UKID87	50.8	-0.7	0.024
Crl_1	54.9	-2.9	UKNW6_078	54.4	-3	0.017
Crl_1	54.9	-2.9	UKNW6_210	54.4	-3	0.017
Crl_1	54.9	-2.9	UKNW6_410	54.7	-3.4	0.049
Crl_1	54.9	-2.9	UKNW9_025	54.6	-3.1	0.016
Crl_1	54.9	-2.9	UKSE6_350	51.3	0.4	0.024
Crl_1	54.9	-2.9	UKSE6_556	51.3	1.1	0.010
Crl_1	54.9	-2.9	UKSE6_618	51.1	0.4	0.003
Crl_1	54.9	-2.9	UKSE6_622	51.1	0.4	0.024
Crl_1	54.9	-2.9	UKSW6_025	50.4	-4.7	0.001
Crl_1	54.9	-2.9	UKSW6_070	50.4	-4.7	0.003

Accession 1	Lat.	Lon.	Accession 2	Lat.	Lon.	p-value
Crl_1	54.9	-2.9	Ullapool4	57.9	-5.1525	0.029
Edburgh_8	55.9681	-3.21833	Hil_1	51	-1.5	0.017
Edburgh_8	55.9681	-3.21833	NFA_8	51.4083	-0.6383	0.017
Edburgh_8	55.9681	-3.21833	UKNW6_078	54.4	-3	0.036
Edburgh_8	55.9681	-3.21833	UKNW6_210	54.4	-3	0.036
Edburgh_8	55.9681	-3.21833	UKNW6_410	54.7	-3.4	0.005
Edburgh_8	55.9681	-3.21833	UKNW9_025	54.6	-3.1	0.010
Edburgh_8	55.9681	-3.21833	UKSW6_070	50.4	-4.7	0.020
Edburgh_8	55.9681	-3.21833	Ullapool4	57.9	-5.1525	0.007
EM_183	51.3	0.5	Ema_1A	51.3	0.5	0.013
EM_183	51.3	0.5	UKID87	50.8	-0.7	0.039
EM_183	51.3	0.5	UKSE6_032	51.3	0.5	0.047
EM_183	51.3	0.5	UKSE6_618	51.1	0.4	0.041
Ema_1A	51.3	0.5	HR_10	51.4083	-0.6383	0.006
Ema_1A	51.3	0.5	PHW_14	51.2878	0.0565	0.011
Ema_1A	51.3	0.5	Sis_1	51.1	0.6	0.041
Ema_1A	51.3	0.5	UKID103	51.8	-0.5	0.017
Ema_1A	51.3	0.5	UKID17	51.4	0.1	0.011
Ema_1A	51.3	0.5	UKID28	52.3	-1.7	0.011
Ema_1A	51.3	0.5	UKID87	50.8	-0.7	0.016
Ema_1A	51.3	0.5	UKNW6_078	54.4	-3	0.025
Ema_1A	51.3	0.5	UKNW6_410	54.7	-3.4	0.024
Ema_1A	51.3	0.5	UKNW9_025	54.6	-3.1	0.025
Ema_1A	51.3	0.5	UKSE6_032	51.3	0.5	0.004
Ema_1A	51.3	0.5	UKSE6_350	51.3	0.4	0.013
Ema_1A	51.3	0.5	UKSE6_544	51.3	1.1	0.026
Ema_1A	51.3	0.5	UKSE6_618	51.1	0.4	0.027
Ema_1A	51.3	0.5	UKSW6_025	50.4	-4.7	0.029
Ema_1A	51.3	0.5	UKSW6_070	50.4	-4.7	0.029
Hil_1	51	-1.5	NFA_8	51.4083	-0.6383	0.024
Hil_1	51	-1.5	PHW_14	51.2878	0.0565	0.039
Hil_1	51	-1.5	Sis_1	51.1	0.6	0.019
Hil_1	51	-1.5	UKID103	51.8	-0.5	0.039
Hil_1	51	-1.5	UKID55	53	-1.1	0.030
Hil_1	51	-1.5	UKNW6_078	54.4	-3	0.029
Hil_1	51	-1.5	UKNW6_210	54.4	-3	0.029
Hil_1	51	-1.5	UKNW6_410	54.7	-3.4	0.012
Hil_1	51	-1.5	UKNW9_025	54.6	-3.1	0.004
Hil_1	51	-1.5	UKSE6_618	51.1	0.4	0.047
Hil_1	51	-1.5	UKSE6_622	51.1	0.4	0.047
Hil_1	51	-1.5	UKSW6_025	50.4	-4.7	0.030
Hil_1	51	-1.5	UKSW6_070	50.4	-4.7	0.030
HR_10	51.4083	-0.6383	NFA_10	51.4083	-0.6383	0.047
HR_10	51.4083	-0.6383	NFA_8	51.4083	-0.6383	0.013
HR_10	51.4083	-0.6383	PHW_14	51.2878	0.0565	0.023
HR_10	51.4083	-0.6383	Sq_1	51.4083	-0.6383	0.047
HR_10	51.4083	-0.6383	UKID103	51.8	-0.5	0.005
HR_10	51.4083	-0.6383	UKID17	51.4	0.1	0.008
HR_10	51.4083	-0.6383	UKID87	50.8	-0.7	0.031
HR_10	51.4083	-0.6383	UKNW6_078	54.4	-3	0.010
HR_10	51.4083	-0.6383	UKNW6_210	54.4	-3	0.029
HR_10	51.4083	-0.6383	UKNW6_410	54.7	-3.4	0.025
HR_10	51.4083	-0.6383	UKNW9_025	54.6	-3.1	0.010
HR_10	51.4083	-0.6383	UKSE6_544	51.3	1.1	0.047
HR_10	51.4083	-0.6383	UKSE6_556	51.3	1.1	0.047
HR_10	51.4083	-0.6383	UKSE6_618	51.1	0.4	0.024
HR_10	51.4083	-0.6383	UKSW6_025	50.4	-4.7	0.020

Accession 1	Lat.	Lon.	Accession 2	Lat.	Lon.	p-value
HR_10	51.4083	-0.6383	UKSW6_070	50.4	-4.7	0.020
HR_10	51.4083	-0.6383	Wis_1	51.3	-0.5	0.043
HR_5	51.4083	-0.6383	NFA_8	51.4083	-0.6383	0.047
HR_5	51.4083	-0.6383	PHW_13	51.2878	0.0565	0.023
HR_5	51.4083	-0.6383	UKID103	51.8	-0.5	0.046
HR_5	51.4083	-0.6383	UKID35	51.3	0.9	0.004
HR_5	51.4083	-0.6383	UKID98	52.3	-1.6	0.016
HR_5	51.4083	-0.6383	UKSE6_350	51.3	0.4	0.006
HR_5	51.4083	-0.6383	UKSE6_556	51.3	1.1	0.003
HR_5	51.4083	-0.6383	UKSW6_157	50.4	-4.7	0.004
Igt_1	51.3	0.3	PHW_13	51.2878	0.0565	0.043
Igt_1	51.3	0.3	UKNW6_197	54.4	-3	0.012
Igt_1	51.3	0.3	UKNW6_355	54.6	-3.1	0.025
NFA_10	51.4083	-0.6383	NFA_8	51.4083	-0.6383	0.047
NFA_10	51.4083	-0.6383	Sq_1	51.4083	-0.6383	0.047
NFA_10	51.4083	-0.6383	UKID103	51.8	-0.5	0.026
NFA_8	51.4083	-0.6383	Sq_1	51.4083	-0.6383	0.013
NFA_8	51.4083	-0.6383	UKID103	51.8	-0.5	0.046
NFA_8	51.4083	-0.6383	UKID35	51.3	0.9	0.039
NFA_8	51.4083	-0.6383	UKID55	53	-1.1	0.036
NFA_8	51.4083	-0.6383	UKNW6_078	54.4	-3	0.002
NFA_8	51.4083	-0.6383	UKNW6_210	54.4	-3	0.003
NFA_8	51.4083	-0.6383	UKNW6_410	54.7	-3.4	0.012
NFA_8	51.4083	-0.6383	UKNW9_025	54.6	-3.1	0.001
NFA_8	51.4083	-0.6383	UKSE6_556	51.3	1.1	0.012
NFA_8	51.4083	-0.6383	UKSE6_618	51.1	0.4	0.024
NFA_8	51.4083	-0.6383	UKSW6_025	50.4	-4.7	0.004
NFA_8	51.4083	-0.6383	UKSW6_070	50.4	-4.7	0.004
NFA_8	51.4083	-0.6383	Ullapool4	57.9	-5.1525	0.015
PHW_13	51.2878	0.0565	Sis_1	51.1	0.6	0.046
PHW_13	51.2878	0.0565	UKID98	52.3	-1.6	0.036
PHW_13	51.2878	0.0565	UKNW6_197	54.4	-3	0.002
PHW_13	51.2878	0.0565	UKNW6_355	54.6	-3.1	0.025
PHW_13	51.2878	0.0565	UKSE6_350	51.3	0.4	0.041
PHW_13	51.2878	0.0565	UKSE6_544	51.3	1.1	0.024
PHW_13	51.2878	0.0565	UKSW6_157	50.4	-4.7	0.004
PHW_13	51.2878	0.0565	UKSW6_329	50.3	-4.8	0.003
PHW_13	51.2878	0.0565	Wis_1	51.3	-0.5	0.049
PHW_14	51.2878	0.0565	Sis_1	51.1	0.6	0.046
PHW_14	51.2878	0.0565	UKID103	51.8	-0.5	0.031
PHW_14	51.2878	0.0565	UKID17	51.4	0.1	0.022
PHW_14	51.2878	0.0565	UKID87	50.8	-0.7	0.024
PHW_14	51.2878	0.0565	UKNW6_410	54.7	-3.4	0.025
PHW_14	51.2878	0.0565	UKSE6_350	51.3	0.4	0.041
PHW_14	51.2878	0.0565	UKSE6_544	51.3	1.1	0.024
PHW_14	51.2878	0.0565	Wis_1	51.3	-0.5	0.024
PHW_22	51.4167	-1.7167	UKID108	52.1	-2.3	0.034
PHW_22	51.4167	-1.7167	UKID64	51.3	1	0.011
PHW_22	51.4167	-1.7167	UKNW6_482	54.4	-2.9	0.004
Poo_1	54.6	-2.8	UKID14	55.2	-2	0.034
Poo_1	54.6	-2.8	UKNW6_425	54.7	-3.4	0.010
Poo_1	54.6	-2.8	UKNW9_010	54.6	-3.1	0.037
Sis_1	51.1	0.6	UKID103	51.8	-0.5	0.016
Sis_1	51.1	0.6	UKID17	51.4	0.1	0.033
Sis_1	51.1	0.6	UKID55	53	-1.1	0.030
Sis_1	51.1	0.6	UKNW6_078	54.4	-3	0.012
Sis_1	51.1	0.6	UKNW6_197	54.4	-3	0.025

Accession 1	Lat.	Lon.	Accession 2	Lat.	Lon.	p-value
Sis_1	51.1	0.6	UKNW6_210	54.4	-3	0.025
Sis_1	51.1	0.6	UKNW6_355	54.6	-3.1	0.024
Sis_1	51.1	0.6	UKNW6_410	54.7	-3.4	0.024
Sis_1	51.1	0.6	UKNW9_025	54.6	-3.1	0.024
Sis_1	51.1	0.6	UKSE6_350	51.3	0.4	0.016
Sis_1	51.1	0.6	UKSE6_544	51.3	1.1	0.026
Sis_1	51.1	0.6	UKSE6_618	51.1	0.4	0.032
Sis_1	51.1	0.6	UKSE6_622	51.1	0.4	0.048
Sis_1	51.1	0.6	UKSW6_025	50.4	-4.7	0.010
Sis_1	51.1	0.6	UKSW6_070	50.4	-4.7	0.029
Sis_1	51.1	0.6	UKSW6_157	50.4	-4.7	0.029
Sis_1	51.1	0.6	UKSW6_329	50.3	-4.8	0.029
Sis_1	51.1	0.6	Wis_1	51.3	-0.5	0.024
Sq_1	51.4083	-0.6383	UKID103	51.8	-0.5	0.046
Sq_1	51.4083	-0.6383	UKNW6_078	54.4	-3	0.029
Sq_1	51.4083	-0.6383	UKNW6_210	54.4	-3	0.029
Sq_1	51.4083	-0.6383	UKNW6_410	54.7	-3.4	0.025
Sq_1	51.4083	-0.6383	UKNW9_025	54.6	-3.1	0.029
Sq_1	51.4083	-0.6383	UKSE6_556	51.3	1.1	0.047
Sq_1	51.4083	-0.6383	UKSW6_025	50.4	-4.7	0.020
Sq_1	51.4083	-0.6383	UKSW6_070	50.4	-4.7	0.020
UKID103	51.8	-0.5	UKID17	51.4	0.1	0.042
UKID103	51.8	-0.5	UKID35	51.3	0.9	0.039
UKID103	51.8	-0.5	UKID87	50.8	-0.7	0.004
UKID103	51.8	-0.5	UKNW6_078	54.4	-3	0.020
UKID103	51.8	-0.5	UKNW9_025	54.6	-3.1	0.029
UKID103	51.8	-0.5	UKSE6_544	51.3	1.1	0.047
UKID103	51.8	-0.5	UKSE6_618	51.1	0.4	0.019
UKID103	51.8	-0.5	UKSW6_070	50.4	-4.7	0.020
UKID108	52.1	-2.3	UKID64	51.3	1	0.030
UKID108	52.1	-2.3	UKNW6_482	54.4	-2.9	0.002
UKID120	56.7333	-5.98333	UKNW6_202	54.4	-3	0.020
UKID120	56.7333	-5.98333	UKSE6_581	51.3	1.1	0.015
UKID14	55.2	-2	UKNW6_425	54.7	-3.4	0.039
UKID14	55.2	-2	UKNW9_010	54.6	-3.1	0.010
UKID17	51.4	0.1	UKID28	52.3	-1.7	0.019
UKID17	51.4	0.1	UKNW6_078	54.4	-3	0.029
UKID17	51.4	0.1	UKNW9_025	54.6	-3.1	0.025
UKID17	51.4	0.1	UKSE6_350	51.3	0.4	0.027
UKID17	51.4	0.1	UKSE6_618	51.1	0.4	0.049
UKID17	51.4	0.1	UKSW6_070	50.4	-4.7	0.029
UKID17	51.4	0.1	Wis_1	51.3	-0.5	0.046
UKID35	51.3	0.9	UKID55	53	-1.1	0.030
UKID35	51.3	0.9	UKID87	50.8	-0.7	0.047
UKID35	51.3	0.9	UKID98	52.3	-1.6	0.007
UKID35	51.3	0.9	UKNW6_078	54.4	-3	0.025
UKID35	51.3	0.9	UKNW6_210	54.4	-3	0.025
UKID35	51.3	0.9	UKNW9_025	54.6	-3.1	0.024
UKID35	51.3	0.9	UKSE6_350	51.3	0.4	0.016
UKID35	51.3	0.9	UKSE6_556	51.3	1.1	0.032
UKID35	51.3	0.9	UKSE6_618	51.1	0.4	0.046
UKID35	51.3	0.9	UKSW6_157	50.4	-4.7	0.012
UKID55	53	-1.1	UKNW6_078	54.4	-3	0.006
UKID55	53	-1.1	UKNW6_210	54.4	-3	0.011
UKID55	53	-1.1	UKNW6_410	54.7	-3.4	0.030
UKID55	53	-1.1	UKNW9_025	54.6	-3.1	0.030
UKID55	53	-1.1	UKSE6_556	51.3	1.1	0.030

Accession 1	Lat.	Lon.	Accession 2	Lat.	Lon.	p-value
UKID55	53	-1.1	UKSE6_618	51.1	0.4	0.030
UKID55	53	-1.1	UKSE6_622	51.1	0.4	0.030
UKID55	53	-1.1	UKSW6_025	50.4	-4.7	0.010
UKID55	53	-1.1	UKSW6_070	50.4	-4.7	0.029
UKID55	53	-1.1	Ullapool4	57.9	-5.1525	0.020
UKID64	51.3	1	UKNW6_482	54.4	-2.9	0.012
UKID87	50.8	-0.7	UKNW6_078	54.4	-3	0.025
UKID87	50.8	-0.7	UKNW9_025	54.6	-3.1	0.024
UKID87	50.8	-0.7	UKSE6_350	51.3	0.4	0.019
UKID87	50.8	-0.7	UKSE6_544	51.3	1.1	0.047
UKID87	50.8	-0.7	UKSE6_556	51.3	1.1	0.047
UKID09	55.6	-3.5	UKNW9_025	54.6	-3.1	0.039
UKID98	52.3	-1.6	UKSE6_556	51.3	1.1	0.030
UKID98	52.3	-1.6	UKSW6_157	50.4	-4.7	0.020
UKNW6_078	54.4	-3	UKNW6_210	54.4	-3	0.013
UKNW6_078	54.4	-3	UKNW6_410	54.7	-3.4	0.010
UKNW6_078	54.4	-3	UKNW9_025	54.6	-3.1	0.027
UKNW6_078	54.4	-3	UKSE6_350	51.3	0.4	0.025
UKNW6_078	54.4	-3	UKSE6_556	51.3	1.1	0.012
UKNW6_078	54.4	-3	UKSE6_618	51.1	0.4	0.004
UKNW6_078	54.4	-3	UKSE6_622	51.1	0.4	0.025
UKNW6_078	54.4	-3	UKSW6_025	50.4	-4.7	0.001
UKNW6_078	54.4	-3	UKSW6_070	50.4	-4.7	0.003
UKNW6_078	54.4	-3	Ullapool4	57.9	-5.1525	0.025
UKNW6_101	54.4	-3	UKNW6_105	54.4	-3	0.013
UKNW6_170	54.6	-3.1	UKNW6_178	54.4	-3	0.027
UKNW6_170	54.6	-3.1	UKNW6_202	54.4	-3	0.041
UKNW6_170	54.6	-3.1	UKSE6_581	51.3	1.1	0.024
UKNW6_178	54.4	-3	UKNW6_202	54.4	-3	0.047
UKNW6_178	54.4	-3	UKSE6_581	51.3	1.1	0.025
UKNW6_197	54.4	-3	UKNW6_355	54.6	-3.1	0.009
UKNW6_197	54.4	-3	UKSE6_544	51.3	1.1	0.004
UKNW6_197	54.4	-3	UKSW6_157	50.4	-4.7	0.024
UKNW6_197	54.4	-3	UKSW6_329	50.3	-4.8	0.003
UKNW6_197	54.4	-3	Ullapool4	57.9	-5.1525	0.025
UKNW6_197	54.4	-3	Wis_1	51.3	-0.5	0.002
UKNW6_202	54.4	-3	UKNW6_306	54.6	-3.1	0.041
UKNW6_202	54.4	-3	UKSE6_565	51.3	1.1	0.025
UKNW6_202	54.4	-3	UKSE6_581	51.3	1.1	0.002
UKNW6_210	54.4	-3	UKNW6_410	54.7	-3.4	0.026
UKNW6_210	54.4	-3	UKNW9_025	54.6	-3.1	0.027
UKNW6_210	54.4	-3	UKSE6_350	51.3	0.4	0.025
UKNW6_210	54.4	-3	UKSE6_556	51.3	1.1	0.004
UKNW6_210	54.4	-3	UKSE6_618	51.1	0.4	0.012
UKNW6_210	54.4	-3	UKSE6_622	51.1	0.4	0.012
UKNW6_210	54.4	-3	UKSW6_025	50.4	-4.7	0.003
UKNW6_210	54.4	-3	UKSW6_070	50.4	-4.7	0.010
UKNW6_210	54.4	-3	Ullapool4	57.9	-5.1525	0.025
UKNW6_306	54.6	-3.1	UKSE6_565	51.3	1.1	0.010
UKNW6_355	54.6	-3.1	UKSE6_544	51.3	1.1	0.010
UKNW6_355	54.6	-3.1	UKSW6_329	50.3	-4.8	0.010
UKNW6_355	54.6	-3.1	Wis_1	51.3	-0.5	0.025
UKNW6_410	54.7	-3.4	UKNW9_025	54.6	-3.1	0.041
UKNW6_410	54.7	-3.4	UKSE6_544	51.3	1.1	0.024
UKNW6_410	54.7	-3.4	UKSE6_556	51.3	1.1	0.024
UKNW6_410	54.7	-3.4	UKSE6_618	51.1	0.4	0.024
UKNW6_410	54.7	-3.4	UKSW6_025	50.4	-4.7	0.003

Accession 1	Lat.	Lon.	Accession 2	Lat.	Lon.	p-value
UKNW6_410	54.7	-3.4	UKSW6_070	50.4	-4.7	0.010
UKNW6_410	54.7	-3.4	Ullapool4	57.9	-5.1525	0.029
UKNW6_425	54.7	-3.4	UKNW9_010	54.6	-3.1	0.009
UKNW9_025	54.6	-3.1	UKSE6_350	51.3	0.4	0.025
UKNW9_025	54.6	-3.1	UKSE6_556	51.3	1.1	0.024
UKNW9_025	54.6	-3.1	UKSE6_618	51.1	0.4	0.024
UKNW9_025	54.6	-3.1	UKSE6_622	51.1	0.4	0.024
UKNW9_025	54.6	-3.1	UKSW6_025	50.4	-4.7	0.010
UKNW9_025	54.6	-3.1	UKSW6_070	50.4	-4.7	0.010
UKNW9_025	54.6	-3.1	Ullapool4	57.9	-5.1525	0.029
UKSE6_032	51.3	0.5	UKSE6_544	51.3	1.1	0.046
UKSE6_032	51.3	0.5	UKSE6_618	51.1	0.4	0.041
UKSE6_192	51.3	0.5	UKSE6_272	51.3	0.4	0.013
UKSE6_254	51.3	0.5	UKSW6_262	50.3	-4.9	0.010
UKSE6_254	51.3	0.5	UKSW6_280	50.3	-4.9	0.003
UKSE6_350	51.3	0.4	UKSE6_544	51.3	1.1	0.033
UKSE6_350	51.3	0.4	UKSE6_556	51.3	1.1	0.033
UKSE6_350	51.3	0.4	UKSE6_618	51.1	0.4	0.027
UKSE6_350	51.3	0.4	UKSE6_622	51.1	0.4	0.041
UKSE6_350	51.3	0.4	UKSW6_025	50.4	-4.7	0.029
UKSE6_350	51.3	0.4	UKSW6_157	50.4	-4.7	0.029
UKSE6_373	51.3	0.4	UKSW6_329	50.3	-4.8	0.029
UKSE6_544	51.3	1.1	UKSW6_157	50.4	-4.7	0.025
UKSE6_544	51.3	1.1	UKSW6_329	50.3	-4.8	0.012
UKSE6_544	51.3	1.1	Wis_1	51.3	-0.5	0.016
UKSE6_556	51.3	1.1	UKSW6_025	50.4	-4.7	0.012
UKSE6_556	51.3	1.1	UKSW6_070	50.4	-4.7	0.025
UKSE6_556	51.3	1.1	UKSW6_157	50.4	-4.7	0.012
UKSE6_618	51.1	0.4	UKSE6_622	51.1	0.4	0.047
UKSE6_618	51.1	0.4	UKSW6_025	50.4	-4.7	0.010
UKSE6_618	51.1	0.4	UKSW6_070	50.4	-4.7	0.010
UKSE6_622	51.1	0.4	UKSW6_025	50.4	-4.7	0.010
UKSW6_025	50.4	-4.7	UKSW6_070	50.4	-4.7	0.013
UKSW6_157	50.4	-4.7	UKSW6_329	50.3	-4.8	0.048
UKSW6_157	50.4	-4.7	Wis_1	51.3	-0.5	0.020
UKSW6_262	50.3	-4.9	UKSW6_280	50.3	-4.9	0.047
UKSW6_329	50.3	-4.8	Wis_1	51.3	-0.5	0.004
Ullapool4	57.9	-5.1525	Wis_1	51.3	-0.5	0.015

A1.2 P<=0.01

Accession 1	Lat.	Lon.	Accession 2	Lat.	Lon.	p-value
02B6	55.9218	-3.17108	Ema_1A	51.3	0.5	0.003
02B6	55.9218	-3.17108	UKID103	51.8	-0.5	0.001
02B6	55.9218	-3.17108	UKID17	51.4	0.1	0.003
12A1	55.8877	-3.16377	UKID103	51.8	-0.5	0.003
12A1	55.8877	-3.16377	UKID87	50.8	-0.7	0.003
12A1	55.8877	-3.16377	UKSE6_350	51.3	0.4	0.001
13B5	55.8858	-3.16015	UKNW9_025	54.6	-3.1	0.010
Cnt_1	51.3	1.1	UKID87	50.8	-0.7	0.003
Cnt_1	51.3	1.1	UKNW6_210	54.4	-3	0.004
Crl_1	54.9	-2.9	NFA_8	51.4083	-0.6383	0.002
Crl_1	54.9	-2.9	UKID55	53	-1.1	0.007
Crl_1	54.9	-2.9	UKSE6_618	51.1	0.4	0.003
Crl_1	54.9	-2.9	UKSW6_025	50.4	-4.7	0.001
Crl_1	54.9	-2.9	UKSW6_070	50.4	-4.7	0.003
Edburgh_8	55.9681	-3.21833	UKNW6_410	54.7	-3.4	0.005
Edburgh_8	55.9681	-3.21833	UKNW9_025	54.6	-3.1	0.010
Edburgh_8	55.9681	-3.21833	Ullapool4	57.9	-5.1525	0.007
Ema_1A	51.3	0.5	HR_10	51.4083	-0.6383	0.006
Ema_1A	51.3	0.5	UKSE6_032	51.3	0.5	0.004
Hil_1	51	-1.5	UKNW9_025	54.6	-3.1	0.004
HR_10	51.4083	-0.6383	UKID103	51.8	-0.5	0.005
HR_10	51.4083	-0.6383	UKID17	51.4	0.1	0.008
HR_5	51.4083	-0.6383	UKID35	51.3	0.9	0.004
HR_5	51.4083	-0.6383	UKSE6_350	51.3	0.4	0.006
HR_5	51.4083	-0.6383	UKSE6_556	51.3	1.1	0.003
HR_5	51.4083	-0.6383	UKSW6_157	50.4	-4.7	0.004
NFA_8	51.4083	-0.6383	UKNW6_078	54.4	-3	0.002
NFA_8	51.4083	-0.6383	UKNW6_210	54.4	-3	0.003
NFA_8	51.4083	-0.6383	UKNW9_025	54.6	-3.1	0.001
NFA_8	51.4083	-0.6383	UKSW6_025	50.4	-4.7	0.004
NFA_8	51.4083	-0.6383	UKSW6_070	50.4	-4.7	0.004
PHW_13	51.2878	0.0565	UKNW6_197	54.4	-3	0.002
PHW_13	51.2878	0.0565	UKSW6_157	50.4	-4.7	0.004
PHW_13	51.2878	0.0565	UKSW6_329	50.3	-4.8	0.003
PHW_22	51.4167	-1.7167	UKNW6_482	54.4	-2.9	0.004
UKID103	51.8	-0.5	UKID87	50.8	-0.7	0.004
UKID108	52.1	-2.3	UKNW6_482	54.4	-2.9	0.002
UKID14	55.2	-2	UKNW9_010	54.6	-3.1	0.010
UKID35	51.3	0.9	UKID98	52.3	-1.6	0.007
UKID55	53	-1.1	UKNW6_078	54.4	-3	0.006
UKNW6_078	54.4	-3	UKSE6_618	51.1	0.4	0.004
UKNW6_078	54.4	-3	UKSW6_025	50.4	-4.7	0.001
UKNW6_078	54.4	-3	UKSW6_070	50.4	-4.7	0.003
UKNW6_197	54.4	-3	UKNW6_355	54.6	-3.1	0.009
UKNW6_197	54.4	-3	UKSE6_544	51.3	1.1	0.004
UKNW6_197	54.4	-3	UKSW6_329	50.3	-4.8	0.003
UKNW6_197	54.4	-3	Wis_1	51.3	-0.5	0.002
UKNW6_202	54.4	-3	UKSE6_581	51.3	1.1	0.002
UKNW6_210	54.4	-3	UKSE6_556	51.3	1.1	0.004
UKNW6_210	54.4	-3	UKSW6_025	50.4	-4.7	0.003
UKNW6_410	54.7	-3.4	UKSW6_025	50.4	-4.7	0.003
UKNW6_425	54.7	-3.4	UKNW9_010	54.6	-3.1	0.009
UKSE6_254	51.3	0.5	UKSW6_280	50.3	-4.9	0.003
UKSW6_329	50.3	-4.8	Wis_1	51.3	-0.5	0.004

APPENDIX 2: GENES -> HAPLOTYPES

Note: Data pertaining to samples gathered from the 'Railway' habitat type are excluded from this and following appendices due to the very small number of samples in this class, which precluded meaningful analysis.

A2.1 HABITAT TYPE: ALL-UK

Gene	HaplotypeIDs
AT1G01160	4
AT1G01860	18
AT1G02340	74
AT1G06010	238
AT1G10650	484
AT1G17070	852
AT1G17060	855
AT1G17180	869
AT1G30000	1624
AT1G30800	1680
AT1G30760	1684
AT1G33200	2000
AT1G33270	1992
AT1G34400	2120
AT1G49630	2976
AT1G50310	3008
AT1G52890	3170
AT1G53070	3182
AT1G54200	3244
AT1G54610	3253
AT1G55020	3264
AT1G63230	3951
AT1G63490	3995
AT1G68910	4537
AT1G77120	5055, 5058
AT1G77610	5083
AT1G78200	5100
AT2G02620	5382
AT2G03960	5500
AT2G12345	6052
AT2G16010	6360
AT2G18800	6550
AT2G19560	6581
AT2G21880	6668
AT2G26240	6897
AT2G27160	6965
AT2G38880	7456
AT2G38970	7462
AT2G40050	7513
AT2G40090	7512
AT3G04150	8082
AT3G05260	8115
AT3G14330	8600
AT3G14475	8643

Gene	HaplotypeIDs
AT3G15940	8727, 8728
AT3G16160	8735
AT3G16840	8757
AT3G18440	8899
AT3G19540	8973
AT3G23955	9289
AT3G26690	9571
AT3G27670	9656
AT3G49710	11733
AT3G49650	11730
AT3G54930	12029
AT3G54920	12028
AT3G56390	12113
AT3G58980	12248
AT3G60390	12321
AT3G61790	12414
AT3G63260	12482
AT3G63470	12492
AT4G00920	12564
AT4G00980	12585
AT4G17060	14800
AT4G18510	14984, 14986
AT4G18500	14990
AT4G19560	15145
AT4G19600	15146
AT4G19700	15155
AT4G19920	15180
AT4G20380	15241
AT4G21810	15338
AT4G32605	15877
AT4G34950	15984
AT4G40000	16278
AT5G05340	16593
AT5G08030	16689
AT5G08620	16701
AT5G11070	16811, 16823
AT5G13550	16912
AT5G14270	16953
AT5G14470	16964
AT5G20490	17354, 17355, 17357
AT5G23910	17572
AT5G25330	17752
AT5G25451	17766, 17776
AT5G27660	17968
AT5G28487	18119
AT5G28469	18120
AT5G38905	19000
AT5G42000	19287
AT5G44005	19594
AT5G45780	19880
AT5G50940	20300
AT5G50900	20297
AT5G53200	20474
AT5G63630	21147
AT5G63990	21169
AT5G64570	21197

Gene	HaplotypeIDs
AT5G64572	21195, 21198
AT5G64685	21216
AT5G66640	21318
AT5G67610	21355

A2.2 HABITAT TYPE: WALL/OUTCROP

Gene	HaplotypeIDs
AT1G01620	33
AT1G01950	26
AT1G05830	215
AT1G06080	237
AT1G06010	238
AT1G21750	1119
AT1G28310	1529
AT1G28307	1522
AT1G30000	1624
AT1G30814	1680
AT1G30760	1684
AT1G31812	1830
AT1G33410	2009
AT1G49340	2966
AT1G52890	3170
AT1G53060	3182
AT1G53140	3184
AT1G54040	3216
AT1G54200	3243, 3244
AT1G54610	3253
AT1G54890	3256
AT1G58248	3445
AT1G60190	3606
AT1G63230	3951
AT1G72310	4788
AT1G76900	5047
AT1G77120	5055, 5058
AT1G77830	5107
AT2G02560	5375, 5376
AT2G21370	6673
AT2G21380	6667
AT2G25980	6891
AT2G26300	6913
AT2G27160	6959, 6965
AT2G28540	7015
AT2G39980	7522
AT2G40090	7512
AT2G40050	7513
AT2G40130	7531
AT2G40180	7535
AT2G40410	7539
AT2G43320	7714
AT3G04340	8085
AT3G14240	8636
AT3G15090	8675
AT3G15110	8677
AT3G16170	8743, 8744

Gene	HaplotypeIDs
AT3G16070	8737
AT3G16720	8779
AT3G16840	8757
AT3G16830	8790
AT3G16785	8734
AT3G18440	8899
AT3G19530	8975
AT3G19540	8973, 8978
AT3G49710	11733
AT3G49650	11730
AT3G54890	12028
AT3G54930	12029
AT4G11660	13952
AT4G11730	13958
AT4G11670	13960
AT4G12070	14036
AT4G15610	14518
AT4G15720	14578
AT4G17100	14812
AT4G17060	14800
AT4G17170	14752
AT4G17360	14818
AT4G17800	14886
AT4G18140	14908
AT4G18510	14984, 14986
AT4G19560	15145
AT4G19600	15146
AT4G20310	15229
AT4G22700	15398
AT4G25900	15586
AT4G29610	15765
AT4G29760	15757
AT4G32230	15869
AT4G32600	15877
AT4G34960	15984
AT5G05340	16593
AT5G08030	16689
AT5G10110	16750
AT5G10960	16806
AT5G11070	16823
AT5G12150	16878
AT5G12400	16874
AT5G12323	16883
AT5G12860	16889
AT5G13560	16912
AT5G13900	16929, 16945
AT5G15120	17003, 17004
AT5G16023	17053
AT5G16010	17044
AT5G16140	17058
AT5G16567	17069
AT5G17960	17168
AT5G18130	17163
AT5G18560	17176
AT5G20490	17354
AT5G23910	17572
AT5G24065	17573

Gene	HaplotypeIDs
AT5G24770	17663
AT5G25451	17776
AT5G26620	17873
AT5G26660	17871
AT5G27480	17946
AT5G27606	17954
AT5G28526	18130
AT5G28523	18125
AT5G43080	19405
AT5G46260	19926
AT5G46540	19939
AT5G49190	20165
AT5G49200	20149
AT5G49770	20211
AT5G53200	20474
AT5G55770	20667
AT5G55760	20668
AT5G55870	20673
AT5G56030	20688
AT5G62420	21070
AT5G62560	21088
AT5G63990	21169
AT5G65110	21229

A2.3 HABITAT TYPE: GARDEN

Gene	HaplotypeIDs
AT1G02340	74
AT1G02400	59
AT1G05250	191
AT1G05360	189
AT1G05740	223
AT1G06010	238
AT1G06670	267
AT1G07090	264
AT1G07280	275
AT1G10650	484
AT1G19630	1001
AT1G21760	1119
AT1G26330	1413
AT1G28310	1522
AT1G28307	1529
AT1G30010	1624
AT1G31820	1842
AT1G31830	1843
AT1G33200	2000, 2001
AT1G33390	2009
AT1G34400	2120
AT1G48095	2880
AT1G51140	3057
AT1G51170	3059
AT1G53340	3140
AT1G53070	3182
AT1G54200	3243

Gene	HaplotypeIDs
AT1G54580	3253
AT1G55230	3257
AT1G54890	3256
AT1G55020	3264
AT1G55580	3279
AT1G55830	3284
AT1G55760	3287
AT1G58242	3443
AT1G60913	3668
AT1G60940	3669
AT1G62990	3959
AT1G63230	3951
AT1G63500	3995
AT1G63820	4022, 4038
AT1G70210	4637
AT1G70209	4642
AT1G74350	4928
AT1G76920	5047
AT1G77120	5055, 5058
AT1G77400	5064
AT1G77310	5070
AT1G77300	5062
AT1G77610	5083
AT1G77885	5078
AT1G77992	5112
AT1G78200	5100
AT1G78380	5113
AT1G78310	5131
AT1G79930	5213
AT1G80590	5223
AT1G80920	5253
AT1G80930	5257
AT2G03960	5500
AT2G11830	5999
AT2G12000	6015
AT2G12340	6035
AT2G17090	6437
AT2G18210	6515
AT2G18810	6548
AT2G20890	6659
AT2G21140	6671
AT2G21385	6673
AT2G25730	6887
AT2G25800	6889
AT2G26190	6905
AT2G26240	6897
AT2G27080	6941
AT2G27090	6954
AT2G28550	6984
AT2G28053	6996
AT2G31550	7160
AT2G31280	7159
AT2G33320	7247
AT2G33880	7250
AT2G34190	7272
AT2G39840	7505
AT2G40050	7513

Gene	HaplotypeIDs
AT2G40090	7512
AT2G40190	7535
AT2G40720	7571
AT2G40830	7562
AT2G41510	7612
AT2G43320	7713, 7714
AT2G43270	7709
AT2G44100	7746
AT3G04510	8100
AT3G06665	8205
AT3G06580	8219
AT3G06483	8220
AT3G09830	8331
AT3G11440	8448
AT3G11580	8486
AT3G11970	8459
AT3G13686	8606
AT3G14330	8600
AT3G15300	8676
AT3G16070	8737
AT3G15990	8736
AT3G16170	8744
AT3G16160	8735
AT3G16720	8779
AT3G16840	8757
AT3G18440	8899
AT3G18550	8901
AT3G27460	9594
AT3G28370	9761
AT3G30705	10179
AT3G43830	10998
AT3G47500	11533
AT3G49710	11733
AT3G49650	11730
AT3G53340	11950
AT3G54930	12029
AT3G56400	12116
AT3G56390	12113
AT3G56600	12125
AT3G57072	12156
AT3G57240	12171
AT3G58120	12216
AT3G58490	12221
AT3G63340	12482
AT4G08073	13449
AT4G08072	13451
AT4G08640	13567
AT4G08820	13578
AT4G11440	13907
AT4G11650	13952
AT4G11670	13960
AT4G12170	14051
AT4G12310	14042
AT4G15730	14586
AT4G16040	14613
AT4G17140	14821
AT4G18490	14995

Gene	HaplotypeIDs
AT4G18465	14991
AT4G18510	14990
AT4G19600	15145, 15146
AT4G19860	15172
AT4G23810	15503
AT4G24840	15497
AT4G25330	15559
AT4G32605	15877
AT4G38990	16213
AT4G39900	16277
AT4G39990	16278
AT4G40010	16266
AT4G39910	16276
AT5G07390	16662
AT5G08030	16689
AT5G08620	16701
AT5G12150	16878
AT5G12323	16879, 16883
AT5G12840	16872
AT5G13240	16888
AT5G12860	16873
AT5G13560	16912
AT5G13900	16929
AT5G14270	16953
AT5G14470	16964
AT5G14730	16967
AT5G14760	16983
AT5G14820	16969
AT5G16140	17058
AT5G16900	17086
AT5G16720	17090
AT5G16680	17091
AT5G17100	17114
AT5G17420	17132
AT5G17430	17128
AT5G20490	17354, 17357
AT5G23170	17518
AT5G23120	17513
AT5G23910	17572
AT5G23970	17584
AT5G24690	17651
AT5G24760	17652
AT5G25360	17742
AT5G25320	17752
AT5G27606	17949
AT5G27660	17962, 17968
AT5G28773	18194
AT5G38905	19000
AT5G43090	19405
AT5G43320	19453
AT5G44005	19594
AT5G44190	19615
AT5G44180	19616
AT5G44565	19669
AT5G45220	19776
AT5G49650	20189
AT5G50900	20297

Gene	HaplotypeIDs
AT5G50940	20300
AT5G51250	20319
AT5G52000	20364
AT5G51870	20368
AT5G53200	20474
AT5G55770	20667
AT5G56770	20734
AT5G59630	20901
AT5G62070	21071
AT5G62670	21102
AT5G63990	21169
AT5G64570	21197, 21198
AT5G64685	21216
AT5G65050	21231
AT5G65180	21210
AT5G65420	21253
AT5G66631	21318
AT5G66817	21335
AT5G67200	21361
AT5G67110	21353
AT5G67610	21355

A2.4 HABITAT TYPE: OTHER

Gene	HaplotypeIDs
AT1G02340	74
AT1G05785	224
AT1G05830	215
AT1G06010	238
AT1G06230	255
AT1G07360	296
AT1G14410	714
AT1G14440	709
AT1G17200	856
AT1G17180	869
AT1G17275	861
AT1G17670	903
AT1G29990	1624
AT1G30100	1614
AT1G30450	1656
AT1G30610	1657
AT1G30810	1684
AT1G30800	1680
AT1G33200	2005
AT1G33190	2000
AT1G33420	2009
AT1G41855	2463
AT1G41850	2462
AT1G48050	2843
AT1G48095	2880
AT1G48440	2904
AT1G50330	3008
AT1G53070	3182
AT1G53460	3204
AT1G53450	3203

Gene	HaplotypeIDs
AT1G55020	3264
AT1G58248	3445
AT1G60200	3606
AT1G63230	3951
AT1G71810	4737
AT1G71870	4753
AT1G71970	4742
AT1G73060	4843
AT1G74480	4935
AT1G75800	4996
AT1G77120	5055, 5058
AT1G77110	5060
AT1G77300	5062
AT1G77310	5070
AT1G77610	5083
AT1G77885	5078
AT1G78380	5113
AT1G78200	5100
AT1G78310	5131
AT2G02560	5375, 5376
AT2G02620	5382
AT2G03960	5500
AT2G07807	5836
AT2G11810	5999
AT2G12345	6035
AT2G14635	6250
AT2G17115	6437
AT2G18735	6498
AT2G18700	6545
AT2G20890	6659
AT2G21880	6668
AT2G22210	6679
AT2G25730	6887
AT2G25820	6889
AT2G26240	6897
AT2G27090	6954
AT2G34170	7272
AT2G40090	7512
AT2G41510	7612
AT2G43410	7714
AT2G43270	7709
AT2G43440	7713
AT2G43690	7725
AT2G43860	7736
AT2G43871	7735
AT2G44950	7780
AT2G44910	7781
AT2G45790	7825
AT3G03300	8021, 8025
AT3G03270	8028
AT3G03530	8037
AT3G05260	8115
AT3G06335	8193
AT3G06660	8205
AT3G08730	8284, 8287
AT3G12120	8473
AT3G13370	8587

Gene	HaplotypeIDs
AT3G13510	8586
AT3G14840	8658
AT3G16160	8735
AT3G16620	8786
AT3G16840	8757
AT3G16830	8734
AT3G18440	8899
AT3G20075	9004
AT3G20705	9027
AT3G20630	9040
AT3G20830	9050
AT3G23620	9263
AT3G27670	9656
AT3G28390	9748
AT3G49480	11724
AT3G49710	11733
AT3G49650	11730
AT3G50170	11761
AT3G54930	12029
AT3G54925	12028
AT3G54910	12026
AT3G55930	12097
AT3G56390	12113
AT3G58980	12248
AT3G59210	12266
AT3G61460	12395
AT3G61790	12414
AT3G63006	12424
AT3G62730	12458
AT3G63470	12492
AT4G00650	12560
AT4G00920	12564
AT4G00990	12585
AT4G00930	12594
AT4G00940	12590
AT4G02460	12758
AT4G02500	12757
AT4G03820	12915, 12916
AT4G04313	12937
AT4G04730	13042
AT4G04880	13083
AT4G08030	13435
AT4G08072	13452
AT4G08076	13449
AT4G09890	13717
AT4G11130	13815
AT4G11385	13910
AT4G11440	13907
AT4G11393	13905
AT4G11660	13952
AT4G14810	14453
AT4G14780	14454
AT4G14970	14476
AT4G15430	14539
AT4G15590	14518
AT4G15560	14558
AT4G19700	15155

Gene	HaplotypeIDs
AT4G21810	15338
AT4G24840	15497
AT4G25330	15559
AT4G25900	15586
AT4G26430	15620
AT4G26620	15617
AT4G26830	15631
AT4G26910	15637
AT4G26790	15635
AT4G27340	15641
AT4G31980	15851
AT4G32605	15877
AT4G32960	15894
AT4G35070	15983
AT4G34950	15984
AT4G39380	16246
AT4G39361	16249
AT4G39890	16277
AT4G40000	16278
AT4G40010	16266
AT5G02290	16336
AT5G03795	16474
AT5G05340	16593
AT5G08030	16689
AT5G08620	16701
AT5G08400	16699
AT5G11060	16811
AT5G11070	16823
AT5G12150	16878
AT5G12323	16883
AT5G13560	16912
AT5G14270	16953
AT5G14470	16964
AT5G14620	16968
AT5G16140	17058
AT5G16270	17071
AT5G16680	17091, 17092
AT5G16715	17090
AT5G16890	17099
AT5G16900	17086, 17093
AT5G17420	17128, 17132
AT5G17430	17129
AT5G18850	17211
AT5G20490	17354, 17355, 17357
AT5G23910	17572
AT5G24060	17573
AT5G25451	17765, 17766, 17776
AT5G26660	17868
AT5G26848	17882
AT5G27410	17947
AT5G27600	17950
AT5G27495	17948
AT5G27660	17968
AT5G28487	18119
AT5G28469	18120

Gene	HaplotypeIDs
AT5G37381	18757
AT5G42000	19287
AT5G44060	19587
AT5G44005	19594
AT5G44190	19615
AT5G44210	19618
AT5G46260	19926
AT5G47750	20027
AT5G48070	20045
AT5G48220	20058
AT5G48100	20046
AT5G48820	20129
AT5G49152	20165
AT5G49130	20162
AT5G50940	20300
AT5G50900	20297
AT5G53200	20474
AT5G59570	20901
AT5G61350	21022
AT5G61240	21025
AT5G61150	21024
AT5G63990	21169
AT5G64430	21190
AT5G64560	21195
AT5G64685	21216
AT5G65305	21246
AT5G65330	21235
AT5G65460	21253
AT5G66170	21264
AT5G65980	21277, 21291
AT5G66140	21305
AT5G66640	21318
AT5G67420	21364
AT5G67630	21355

APPENDIX 3: HAPLOTYPES -> SAMPLES

The data presented in this appendix represent a small sample of the haplotypes marked by SelectionFinder analysis – specifically, those corresponding to LRR-type genes, or those known to be associated with flowering time. The full set of this data is provided as supplementary data alongside the electronic version of this document.

A3.1 HABITAT TYPE: ALL-UK

Haplotype 18

Chromosome 1: 143704-444206

Window	Start	End	Accessions
6	143704	167692	UKID101
7	169429	198389	UKID101
8	198692	228318	EM_183 UKID101 UKID57 UKNW6_079 UKSE6_640 HR_5 Cnt_1 UKSE6_618 UKID35 HR_10 Ema_1A UKID28 UKID17 UKID87 02B6 UKID103 12A1 UKSE6_624
9	228432	250479	Sis_1 EM_183 UKID101 UKNW6_101 UKID108 PHW_22 UKNW6_482 UKID64 UKID57 UKNW6_079 PHW_26 UKNW6_019 Edi_0 UKSE6_640 HR_5 Cnt_1 UKSE6_618 UKID35 HR_10 Ema_1A UKID28 Edburgh_5 UKID17 UKID87 UKNW6_259 02B6 UKID103 UKNW6_105 12A1 UKSE6_624
10	250624	274180	EM_183 UKID101 UKNW6_101 UKID108 PHW_31 PHW_22 PHW_10 UKNW6_482 UKID64 UKID109 PHW_14 UKNW6_460 CIBC_5A 13B5 UKSW6_202 UKSE6_351 EM_134 UKID57 Ullapool3 UKNW9_010 UKSE6_597 UKSE6_350 Boot_1 UKNW6_079 UKSW6_070 Poo_1 PHW_26 UKNW6_019 Edi_0 UKNW6_078 UKSE6_640 HR_5 Edburgh_8 UKSW6_227 UKNW9_025 Crl_1 Cnt_1 UKSE6_618 UKID35 Sq_1 HR_10 Ema_1A NFA_8 UKID28 Edburgh_5 Hil_1 UKID17 UKID87 UKNW6_259 02B6 UKID103 UKNW6_105 12A1 UKSE6_624
11	274842	304281	UKID101 UKID108 PHW_22 UKNW6_482 UKID64 UKID57 Ullapool3 UKNW9_010 UKSE6_597 UKSE6_350 Boot_1 UKNW6_079 UKSW6_070 Poo_1 PHW_26 UKNW6_019 Edi_0 UKNW6_078 UKSE6_640 HR_5 Edburgh_8 UKSW6_227 UKNW9_025 Crl_1 Cnt_1 UKSE6_618 UKID35 Sq_1 HR_10 Ema_1A NFA_8 UKID28 Edburgh_5 Hil_1 UKID17 UKID87 UKNW6_259 02B6 UKID103 UKNW6_105 12A1 UKSE6_624
12	304565	346038	UKID101 UKID57 UKNW6_079 PHW_26 UKNW6_019 Edi_0 UKNW6_078 UKSE6_640 HR_5 Edburgh_8 UKSW6_227 UKNW9_025 Crl_1 Cnt_1 UKSE6_618 UKID35 Sq_1 HR_10 Ema_1A NFA_8 UKID28 Edburgh_5 Hil_1 UKID17 UKID87 UKNW6_259 02B6 UKID103 UKNW6_105 12A1 UKSE6_624 UKSE6_622 Igt_1 UKID33 UKID98 Lc_0 NFC_20 Sq_8 UKID65 UKID109 PHW_14 Kyl_1 UKSE6_311 For_2 UKSW6_262 CIBC_2 UKSE6_032 UKID15 Kil_0 PHW_31

Kent UKNW6_460 CIBC_4 UKID34 UKNW6_425 UKSE6_278 Set_1 NFA_10 UKSW6_220 UKSW6_280 UKID58 Edi_1 CIBC_5B UKID54 CIBC_17 UKSE6_254

13 348970 379604 UKID101 UKID57 UKNW6_079 UKNW6_078 UKSE6_640 HR_5 Edburgh_8 UKSW6_227 UKNW9_025 Crl_1 Cnt_1 UKSE6_618 UKID35 Sq_1 HR_10 Ema_1A NFA_8 UKID28 Edburgh_5 Hil_1 UKID17 UKID87 UKNW6_259 02B6 UKID103 UKNW6_105 12A1 UKSE6_624 PHW_13 UKSE6_622 UKSE6_544 UKID65 Ullapool4 Sis_1 UKSW6_329 UKNW6_197 UKSW6_157 NFA_10 Wis_1 UKNW6_101 UKSE6_350 UKSW6_070

14 381773 408649 UKSE6_640 HR_5 Cnt_1 UKSE6_618 UKID35 HR_10 Ema_1A UKID28 UKID17 UKID87 02B6 UKID103 12A1 UKSE6_624 Sis_1 UKSE6_032

15 409707 427813 UKSE6_032

16 428014 444206 UKSE6_032

==========

Haplotype 3264

Chromosome 1: 20503582-20555114

Window Start End Accessions

670 20503582 20555114 UKNW6_210 UKID33 Igt_1 UKID80 UKSW6_025 UKID87 UKID28 PHW_14 UKNW6_418 PHW_20 Ullapool4 UKID101 UKSW6_070 UKID17 Cnt_1 UKID109 UKNW6_078 UKID72 UKNW6_019 Sq_8 Wis_1 UKID48 12A1 UKID55 Sis_1 HR_10 Crl_1 UKNW6_460 UKSW6_337 NFA_10 Ema_1A UKNW6_410 UKSE6_032 UKSE6_544 For_2

==========

Haplotype 8643

Chromosome 3: 4846991-4887202

Window Start End Accessions

1791 4846991 4864392 09A3 Asp_1 For_2 HR_10 Igt_1 Kil_0 PHW_31 Sq_8 Ty_0 UKID101 UKID113 UKID15 UKID33 UKID34 UKID35 UKID39 UKID48 UKID58 UKID80 UKID98 UKNW6_170 UKNW6_259 UKNW6_306 UKNW6_418 UKSE6_278 UKSE6_349 UKSE6_351 UKSE6_565 UKSE6_597 UKSE6_640 UKSW6_227

1792 4864456 4887202 Igt_1 PHW_31 Sq_8 UKID113 UKID33 UKID35 UKID39 UKID98 UKNW6_418 UKSE6_278 UKSE6_351 UKSE6_597 UKSE6_640 UKSW6_227 CIBC_2

==========

Haplotype 11733

Chromosome 3: 18383727-18484667

Window Start End Accessions

2328 18383727 18405985 02B6 12A1 13B5 Asp_1 CIBC_5A Cnt_1 Crl_1 Edburgh_8 Ema_1A For_2 Hil_1 HR_10 NFA_10 NFA_8 NFC_20 PHW_14 UKID101 UKID103 UKID17 UKID55 UKID80 UKID87 UKID09 UKNW6_078 UKNW6_079 UKNW6_101 UKNW6_105 UKNW6_210 UKNW6_410 UKNW6_436 UKNW9_025 UKSE6_192 UKSE6_272 UKSE6_373 UKSE6_618 UKSE6_622 UKSE6_624 UKSE6_626 UKSW6_025 UKSW6_070 Wis_1

2329 18407136 18453900 02B6 12A1 13B5 Asp_1 CIBC_5A Cnt_1 Crl_1 Edburgh_8 Ema_1A Hil_1 HR_10 NFA_10 NFA_8 PHW_14 UKID103 UKID17 UKID55 UKID87 UKID09 UKNW6_078 UKNW6_079 UKNW6_101 UKNW6_105 UKNW6_210 UKNW6_410 UKNW6_436 UKNW9_025 UKSE6_373 UKSE6_618 UKSE6_622 UKSE6_624 UKSE6_626 UKSW6_025 UKSW6_070 Wis_1 Sis_1 UKNW6_355 PHW_13 EM_183 UKNW6_197 UKSE6_544 UKSE6_032 Ullapool4 Igt_1 UKSW6_329

2330 18454342 18484667 02B6 12A1 13B5 Asp_1 CIBC_5A Cnt_1 Crl_1 Edburgh_8 Ema_1A Hil_1 HR_10 NFA_10 NFA_8 PHW_14 UKID103 UKID17 UKID55 UKID87 UKID09 UKNW6_078 UKNW6_079 UKNW6_101 UKNW6_105 UKNW6_210 UKNW6_410 UKNW6_436 UKNW9_025 UKSE6_373 UKSE6_618 UKSE6_622 UKSE6_624 UKSE6_626 UKSW6_025 UKSW6_070 Wis_1 Sis_1 EM_183 UKNW6_040 UKSE6_272 PHW_28 UKSE6_192 Coc_1 UKSW6_227 Ba_1 UKID39 UKID101 CIBC_17 Ema_1B EM_134 Mc_0 UKID15 09A3 PHW_10

==========

Haplotype 12113

Chromosome 3: 20837697-20940708

Window Start End Accessions

2413 20837697 20879578 EM_183

2414 20882525 20912095 EM_183 Kyl_1 NFA_10 Wis_1 UKID35 UKNW6_386 CIBC_4 PHW_28 UKID55 Alst_1 Poo_1 UKID98 Sis_1 UKSE6_556 UKID09 Ty_0 PHW_31 UKNW6_482 UKNW6_355 UKSE6_640 UKSW6_227 UKNW6_105 UKSW6_157 Ba_1 PHW_13 UKNW6_436 Edburgh_5 Ema_1A 02B6 UKSE6_626 UKNW6_425 UKSE6_618 UKSE6_350 CIBC_5B UKNW6_197 NFC_20 UKNW9_010 UKID101 Sq_8 Hil_1 UKSE6_597 Edburgh_8 Boot_1 UKSE6_544 UKNW6_259 UKID48 PHW_14 HR_5 UKID17 UKID34 UKID65 UKSE6_624 UKID103 UKNW6_101 Ullapool4 UKID109 Ullapool3 Igt_1 UKID58 UKID14 HR_10 UKSW6_329

2415 20912771 20940708 NFA_10 Wis_1 UKID35 CIBC_4 UKID55 Alst_1 Poo_1 UKID09 PHW_31 UKNW6_482 UKNW6_355 UKSE6_640 UKSW6_227 UKNW6_105 UKNW6_436 Edburgh_5 Ema_1A 02B6 UKSE6_626 UKNW6_425 UKSE6_618 CIBC_5B UKNW6_197 NFC_20 UKNW9_010 Hil_1 UKSE6_597 Edburgh_8 Boot_1 UKSE6_544 UKID48 PHW_14 UKID17 UKID65 UKSE6_624 UKID103 Ullapool4 UKID109 Ullapool3 Igt_1 UKID58 UKID14 HR_10 UKSW6_329

==========

Haplotype 15145

Chromosome 4: 10654632-10696635

Window Start End Accessions

2983 10654632 10664008 09A3 CIBC_2 CIBC_4 CIBC_5B Edburgh_8 Frd_1 Hil_1 Kil_0 Kyl_1 PHW_13 Sis_1 Sq_1 UKID113 UKID120 UKID64 UKNW6_101 UKNW6_105 UKNW6_197 UKNW6_259 UKNW6_410 UKSE6_373 UKSE6_544 UKSE6_597 UKSW6_329

2984 10664083 10677188 09A3 CIBC_2 CIBC_5B Edburgh_8 Frd_1 Hil_1 Kil_0 Kyl_1 PHW_13 Sis_1 UKID120 UKID64 UKNW6_197 UKNW6_259 UKNW6_410 UKSE6_544 UKSE6_597 UKSW6_329 UKNW6_040 UKID55 UKID98 UKSW6_220 Ty_0 CIBC_5A EM_134 UKID72 UKID33 UKID65 PHW_10

2985 10677370 10696635 09A3 CIBC_2 CIBC_5B Edburgh_8 Hil_1 PHW_13 Sis_1 UKNW6_197 UKNW6_259 UKNW6_410 UKSE6_544 UKSE6_597 UKSW6_329 UKSE6_315 Poo_1 UKNW6_355 UKNW6_425 UKNW9_010 Ema_1B Boot_1 UKID14

==========

Haplotype 15241

Chromosome 4: 10989483-11032245

Window Start End Accessions

3001 10989483 11005960 02B6 12A1 13B5 CIBC_2 CIBC_5B Cnt_1 Edburgh_5 Edburgh_8 EM_183 Ema_1A Hil_1 HR_10 Igt_1 Kil_0 Lc_0 NFA_10 PHW_13 PHW_14 PHW_20 Sis_1 UKID101 UKID103 UKID17 UKID28 UKID39 UKID48 UKID57 UKID80 UKID87 UKNW6_079 UKNW6_101 UKNW6_105 UKNW6_197 UKNW6_259 UKNW6_386 UKNW6_410 UKNW6_436 UKSE6_032 UKSE6_350 UKSE6_544 UKSE6_565 UKSE6_597 UKSE6_626 UKSW6_220 UKSW6_329 Wis_1

3002 11006496 11019320 02B6 12A1 13B5 CIBC_2 CIBC_5B Cnt_1 Edburgh_5 EM_183 Ema_1A HR_10 Lc_0 PHW_13 PHW_14 Sis_1 UKID101 UKID103 UKID17 UKID28 UKID39 UKID48 UKID57 UKID87 UKNW6_079 UKNW6_101 UKNW6_105 UKNW6_197 UKNW6_436 UKSE6_350 UKSE6_544 UKSE6_597 UKSE6_626 UKSW6_329 Wis_1 UKNW6_019

3003 11019809 11032245 02B6 12A1 13B5 CIBC_2 CIBC_5B Cnt_1 Edburgh_5 Ema_1A HR_10 Lc_0 PHW_13 PHW_14 Sis_1 UKID101 UKID103 UKID17 UKID28 UKID39 UKID48 UKID57 UKID87 UKNW6_079 UKNW6_101 UKNW6_105 UKNW6_197 UKNW6_436 UKSE6_350 UKSE6_544 UKSE6_597 UKSE6_626 UKSW6_329 Wis_1

==========

Haplotype 16953

Chromosome 5: 4574680-4632244

Window Start End Accessions

3416 4574680 4607113 13B5 Cnt_1 Crl_1 HR_10 NFA_8 UKID35 UKID48 UKID55 UKNW6_078 UKNW6_210 UKNW6_410 UKSE6_556 UKSE6_640 UKSW6_025 UKSW6_070 UKSW6_337

3417 4607453 4632244 13B5 Cnt_1 Crl_1 HR_10 NFA_8 UKID35 UKID48 UKID55 UKNW6_078 UKNW6_210 UKNW6_410 UKSE6_556 UKSE6_640 UKSW6_025 UKSW6_070 UKSW6_337 UKSE6_622 UKNW9_025 Sis_1 UKID09 Coc_1 Ema_1A UKID28 02B6 UKSE6_618 UKSE6_350 EM_183 Hil_1 UKSE6_373 UKSE6_032 PHW_14 UKID17 UKID65 UKNW6_101 UKID103

==========

Haplotype 17357

Chromosome 5: 6911867-6953711

Window Start End Accessions

3512 6911867 6935750 Wis_1 UKID55 Edburgh_5 UKNW6_410 Alst_1 UKID101 13B5 UKID103 UKSW6_025 Sq_1 UKSE6_350 UKNW6_210 UKSE6_556 UKID54 PHW_20 UKNW6_078 Kyl_1 UKSW6_337 NFA_10 UKID72 02B6 PHW_14 UKID87 UKSE6_618 Crl_1 CIBC_5A HR_10 NFA_8 Sis_1 Ema_1A Cnt_1 UKID17 Coc_1

3513 6935945 6953711 Wis_1 UKID55 Edburgh_5 UKNW6_410 Alst_1 UKID101 13B5 UKID103 UKSW6_025 Sq_1 UKNW6_210 UKSE6_556 UKID54 PHW_20 UKNW6_078 Kyl_1 UKSW6_337 NFA_10 UKID72 02B6 PHW_14 UKID87 UKSE6_618 Crl_1 CIBC_5A HR_10 NFA_8 Sis_1 Ema_1A Cnt_1 UKID17 Coc_1 12A1 UKID15 UKSW6_070

==========

Haplotype 17354

Chromosome 5: 6911867-6953711

Window Start End Accessions

3512 6911867 6935750 09A3 12A1 Ba_1 Boot_1 Ema_1B For_2 Frd_1 Hil_1 HR_5 Kil_0 Lc_0 PHW_22 PHW_31 Poo_1 UKID108 UKID14 UKID35 UKID39 UKID48 UKID57 UKID64 UKID65 UKID09 UKID98 UKNW6_019 UKNW6_040 UKNW6_101 UKNW6_105 UKNW6_259 UKNW6_386 UKNW6_425 UKNW6_460 UKNW9_010 UKNW9_025 UKSE6_315 UKSE6_349 UKSE6_622 UKSE6_624 UKSE6_626 UKSW6_202 UKSW6_220

3513 6935945 6953711 09A3 Ba_1 Boot_1 Ema_1B For_2 Frd_1 Hil_1 HR_5 Kil_0 Lc_0 PHW_22 PHW_31 Poo_1 UKID108 UKID14 UKID35 UKID39 UKID48 UKID57 UKID64 UKID65 UKID09 UKID98 UKNW6_019 UKNW6_040 UKNW6_101 UKNW6_105 UKNW6_259 UKNW6_386 UKNW6_425 UKNW6_460 UKNW9_010 UKNW9_025 UKSE6_315 UKSE6_349 UKSE6_622 UKSE6_624 UKSE6_626 UKSW6_202 UKSW6_220 UKSE6_581 UKNW6_482 UKNW6_436 UKNW6_202 UKSE6_311 UKSE6_351 UKSE6_373 UKID33 Ullapool3

==========

Haplotype 17355

Chromosome 5: 6911867-6953711

Window Start End Accessions

3512 6911867 6935750 CIBC_17 CIBC_2 CIBC_4 CIBC_5B Edi_0 Edburgh_8 Igt_1 Kent NFC_20 PHW_13 PHW_28 UKID109 UKID113 UKID28 UKID34 UKID58 UKNW6_050 UKNW6_079 UKNW6_197 UKNW6_355 UKNW6_418 UKSE6_192 UKSE6_272 UKSE6_278 UKSE6_544 UKSE6_640 UKSW6_227 UKSW6_329 Ullapool4

3513 6935945 6953711 CIBC_17 CIBC_2 CIBC_4 CIBC_5B Edi_0 Edburgh_8 Igt_1 Kent NFC_20 PHW_13 PHW_28 UKID109 UKID113 UKID28 UKID34 UKID58 UKNW6_050 UKNW6_079 UKNW6_197 UKNW6_355 UKNW6_418 UKSE6_192 UKSE6_272 UKSE6_278 UKSE6_544 UKSE6_640 UKSW6_227 UKSW6_329 Ullapool4 Asp_1

==========

Haplotype 19880

Chromosome 5: 18561437-18618414

Window Start End Accessions

4010 18561437 18571339 UKID87 UKID98 PHW_28 CIBC_5A UKID17 NFA_8 Igt_1 UKSE6_618 UKID35 PHW_14 Edburgh_8 UKID58 UKNW6_197 EM_183 UKSW6_025 UKID48 UKSW6_157 UKNW9_025 UKSE6_350 Ema_1A HR_5 UKSW6_329 UKSE6_622 PHW_13 Wis_1 12A1 NFA_10 09A3 UKID55 UKID113 UKID39 UKSE6_032 HR_10 UKSW6_070 UKID103 Ullapool4 Sis_1 UKSE6_544 02B6 UKSE6_626 Sq_1 Hil_1 CIBC_4 UKNW6_210 13B5 UKSW6_227 Crl_1 UKSE6_640 UKNW6_410 Cnt_1 Alst_1 UKNW6_078 UKNW6_355 UKSE6_556 CIBC_2

4011 18574247 18583234 UKSE6_597 UKSE6_032

4012 18583305 18600057 NFC_20 Ba_1 UKSE6_032

4013 18600718 18618414 Ba_1

==========

Haplotype 21318

Chromosome 5: 26546023-26645553

Window Start End Accessions

4314 26546023 26566676 Edburgh_5 UKNW6_482 UKID65 For_2 UKNW6_105 UKID14 UKSE6_556 Ullapool4 Frd_1 UKSW6_337 UKID64 Cnt_1 PHW_13 Ema_1A UKID108 09A3 Igt_1 UKID87 UKNW6_425 UKSE6_624 UKNW6_078 UKSE6_618 UKID48 UKID98 UKSE6_597 UKNW6_259 UKID109 PHW_22 UKID15 Boot_1 UKSW6_025 UKID103 Sis_1 UKNW6_019 Edi_0 Ullapool3 Poo_1 HR_5 UKNW6_101 UKNW6_306 PHW_10 UKSE6_350 UKNW6_355 UKNW6_197 Alst_1 UKNW6_418 UKID35 UKID28 UKSE6_622 UKNW6_410 UKNW9_025 UKNW9_010 UKNW6_210 UKID55 Wis_1 Crl_1 02B6 UKSE6_351 NFA_8

4315 26566823 26586649 Edburgh_5 UKNW6_482 UKID65 For_2 UKNW6_105 UKID14 UKSE6_556 Ullapool4 Frd_1 UKSW6_337 UKID64 Cnt_1 PHW_13 Ema_1A UKID108 09A3 Igt_1 UKID87 UKNW6_425 UKSE6_624 UKNW6_078 UKSE6_618 UKID48 UKID98 UKSE6_597 UKNW6_259 UKID109 PHW_22 UKID15 Boot_1 UKSW6_025 UKID103 Sis_1 UKNW6_019 Edi_0 Ullapool3 Poo_1 HR_5 UKNW6_101 UKNW6_306 PHW_10 UKSE6_350 UKNW6_355 UKNW6_197 Alst_1 UKNW6_418 UKID35 UKID28 UKSE6_622 UKNW6_410 UKNW9_025 UKNW9_010 UKNW6_210 UKID55 Wis_1 Crl_1 02B6 UKSE6_351 NFA_8

4316 26586679 26606805 Edburgh_5 UKNW6_482 UKID65 UKID17 For_2 UKNW6_105 UKID14 UKSE6_556 Ullapool4 Frd_1 UKSW6_337 UKSE6_640 UKSW6_329 UKID64 12A1 UKSE6_373 Cnt_1 PHW_13 Ema_1A UKID108 UKID34 09A3 UKNW6_050 Igt_1 UKID58 UKID87 UKNW6_425 UKSE6_624 UKNW6_078 UKSE6_618 UKID48 UKID98 UKSE6_597 UKNW6_259 UKID109 PHW_22 UKID15 UKSW6_157 Boot_1 UKSW6_025 UKID103 Sis_1 PHW_20 UKNW6_019 NFA_10 Edi_0 Ullapool3 Sq_1 Poo_1 HR_5 UKNW6_101 UKNW6_306 PHW_10 UKSE6_350 UKNW6_355 UKNW6_197 Alst_1 UKNW6_418 UKID35 UKID54 UKID28 UKSE6_622 HR_10 UKNW6_410 UKNW9_025 UKNW9_010 UKNW6_210 UKID55 Wis_1 Crl_1 02B6 UKSE6_351 NFA_8

4317 26607116 26628408 Edburgh_5 UKNW6_482 UKID65 UKID17 For_2 UKNW6_105 UKID14 UKSE6_556 Ullapool4 Frd_1 UKSW6_337 UKSE6_640 UKSW6_329 UKID64 UKSE6_373 Cnt_1 PHW_13 Ema_1A UKID108 UKID34 09A3 UKNW6_050 Igt_1 UKID58 UKID87 UKNW6_425 UKSE6_624 Ema_1B UKNW6_078 UKSE6_618 UKID48 UKID98 UKSE6_597 UKNW6_259 UKID109 PHW_22 UKID15 UKSW6_157 Boot_1 UKSW6_025 UKID103 Sis_1 PHW_20 UKNW6_019 NFA_10 Edi_0 Ullapool3 Sq_1 Poo_1 HR_5 UKNW6_101 UKNW6_306 PHW_10 UKSE6_350 UKNW6_355 UKNW6_197 Alst_1 UKNW6_418 PHW_26 UKID35 UKID54 UKID28 UKSE6_622 HR_10 UKNW6_410 UKNW9_025 UKNW9_010 UKNW6_210 UKID55 Wis_1 Crl_1 02B6 UKSE6_351 NFA_8

4318 26628440 26645553 UKNW6_482 UKID65 UKID17 For_2 UKNW6_105 UKID14 UKSE6_556 UKSE6_640 UKSW6_329 UKID64 UKSE6_373 Cnt_1 PHW_13 Ema_1A UKID108 UKID34 09A3 UKNW6_050 UKID58 UKID87 UKNW6_425 UKSE6_624 Ema_1B UKNW6_078 UKSE6_618 UKID48 UKID98 UKNW6_259 UKID109 PHW_22 UKID15 Boot_1 UKSW6_025 UKID103 Sis_1 PHW_20 UKNW6_019 NFA_10 Edi_0 Sq_1 Poo_1 HR_5 UKNW6_101 UKNW6_306 UKSE6_350 Alst_1 UKNW6_418 PHW_26 UKID35 UKID54 UKSE6_622 HR_10 UKNW6_410 UKNW9_025 UKNW9_010 UKNW6_210 UKID55 12A1 Crl_1 02B6 NFA_8

==========

A3.2 HABITAT TYPE: WALL/ROCKY OUTCROP

Haplotype 1119

Chromosome 1: 7590135-7705277

Window Start End Accessions

263 7590135 7636992 Ba_1 Kil_0 Set_1 UKID57 UKNW6_050 UKNW6_101 UKNW6_105 UKSE6_254 UKSE6_311 UKSE6_315 UKSE6_544 UKSE6_597

264 7637374 7657534 Ba_1 Kil_0 Set_1 UKID57 UKNW6_050 UKNW6_101 UKNW6_105 UKSE6_254 UKSE6_311 UKSE6_315 UKSE6_544 Kyl_1 UKSE6_581 UKNW6_178 UKNW6_079

265 7658135 7705277 Ba_1 UKID57 UKSE6_544 UKNW6_079 UKNW6_078 UKSE6_556 UKNW6_210 UKSE6_597 UKSW6_025

==========

Haplotype 4788

Chromosome 1: 27168100-27285159

Window Start End Accessions

950 27168100 27206834 Lc_0 UKID80

951 27208549 27234249 Ty_0 UKNW6_105 UKSE6_254 UKSE6_581 For_2 UKNW6_101 Ba_1 Mc_0 UKNW6_259 UKSE6_272 UKSE6_315 Lc_0 UKID80

952 27236199 27285159 UKNW6_105 UKNW6_101 UKNW6_306 UKNW6_259

==========

Haplotype 8085

Chromosome 3: 1131386-1187823

Window Start End Accessions

1668 1131386 1159608 UKNW6_170 UKSE6_544 UKNW6_178 Ba_1 UKNW6_105 UKSE6_311 UKSE6_272 Lc_0

1669 1159812 1187823 UKSE6_544 UKNW6_105 Lc_0

==========

Haplotype 11733

Chromosome 3: 18383727-18484667

Window Start End Accessions

2328 18383727 18405985 For_2 UKID80 UKNW6_078 UKNW6_079 UKNW6_101 UKNW6_105 UKNW6_210 UKSE6_272 UKSW6_025

2329 18407136 18453900 UKNW6_078 UKNW6_079 UKNW6_101 UKNW6_105 UKNW6_210 UKSW6_025 UKSE6_544

2330 18454342 18484667 UKNW6_078 UKNW6_079 UKNW6_101 UKNW6_105 UKNW6_210 UKSW6_025 UKNW6_040 UKSE6_272 UKSW6_227 Ba_1 Mc_0

==========

Haplotype 14886

Chromosome 4: 9871771-9929254

Window Start End Accessions

2943 9871771 9897344 UKSE6_315 UKSE6_544

2944 9899335 9929254 UKSE6_544 UKNW6_078 Frd_1 UKNW6_019 UKNW6_050 UKSW6_025 UKNW6_079

==========

Haplotype 15145

Chromosome 4: 10654632-10696635

Window Start End Accessions

2983 10654632 10664008 Frd_1 Kil_0 Kyl_1 UKID120 UKNW6_101 UKNW6_105 UKNW6_259 UKSE6_544 UKSE6_597

2984 10664083 10677188 Frd_1 Kil_0 Kyl_1 UKID120 UKNW6_259 UKSE6_544 UKSE6_597 UKNW6_040 Ty_0

2985 10677370 10696635 UKNW6_259 UKSE6_544 UKSE6_597 UKSE6_315

==========

Haplotype 15229

Chromosome 4: 10948771-10978769

Window Start End Accessions

2999 10948771 10978769 UKNW6_079 UKNW6_101 Frd_1 UKNW6_040 UKSE6_311 UKSE6_544 UKSW6_227 UKSE6_315 UKSE6_556 UKSE6_597 UKNW6_105

==========

Haplotype 17003

Chromosome 5: 4874569-4942411

Window Start End Accessions

3426 4874569 4900675 Kyl_1 Mc_0 Ty_0 UKID120 UKNW6_170 UKNW6_178 UKSE6_565 UKSE6_581

3427 4905205 4942411 Mc_0 Ty_0 UKID120 UKNW6_170 UKNW6_178 UKSE6_565 UKSE6_581 UKID108 UKNW6_078 UKSE6_556 UKNW6_210 UKSW6_025

==========

Haplotype 17004

Chromosome 5: 4874569-4942411

Window Start End Accessions

3426 4874569 4900675 Frd_1 UKID108 UKID57 UKNW6_079 UKNW6_259 UKSE6_272 UKSE6_278 UKSE6_311 UKSE6_315

3427 4905205 4942411 Frd_1 UKNW6_079 UKSE6_272 UKSE6_278 UKSE6_311 UKSE6_315 Lc_0

==========

Haplotype 17053

Chromosome 5: 5195272-5257336

Window Start End Accessions

3436 5195272 5218166 Mc_0 UKID120 UKID80 UKNW6_170 UKNW6_178 UKNW6_259 UKSE6_565 UKSE6_581

3437 5218307 5234988 Mc_0 UKNW6_170 UKSE6_581 UKSE6_315 UKSE6_272 Ty_0 Set_1

3438 5238171 5257336 Mc_0 UKNW6_170 UKSE6_581 UKSE6_315 UKSE6_272 Ty_0 UKNW6_040 UKSE6_278 UKID108 Frd_1 UKSE6_311 UKNW6_050 UKSE6_565 Kil_0 UKSE6_544 UKNW6_178 UKNW6_101 UKID120 UKID80 For_2 UKNW6_079

==========

Haplotype 17044

Chromosome 5: 5218307-5257336

Window Start End Accessions

3437 5218307 5234988 UKSW6_227 UKNW6_105 UKSE6_311 UKID108 UKNW6_079 UKSE6_565 UKNW6_101

3438 5238171 5257336 Set_1

==========

Haplotype 17354

Chromosome 5: 6911867-6953711

Window Start End Accessions

3512 6911867 6935750 Ba_1 For_2 Frd_1 Kil_0 Lc_0 UKID108 UKID57 UKNW6_019 UKNW6_040 UKNW6_101 UKNW6_105 UKNW6_259 UKSE6_315

3513 6935945 6953711 Ba_1 For_2 Frd_1 Kil_0 Lc_0 UKID108 UKID57 UKNW6_019 UKNW6_040 UKNW6_101 UKNW6_105 UKNW6_259 UKSE6_315 UKSE6_581 UKSE6_311

==========

Haplotype 19926

Chromosome 5: 18740583-18779109

Window Start End Accessions

4021 18740583 18759801 UKID108 UKNW6_078 UKNW6_079 UKNW6_101 UKNW6_105 UKNW6_210 UKSE6_544 UKSE6_556 UKSW6_025 UKSW6_227

4022 18760126 18779109 UKID108 UKNW6_078 UKNW6_079 UKNW6_210 UKSE6_544 UKSE6_556 UKSW6_025 UKSW6_227 UKSE6_597 Set_1

==========

Haplotype 19939

Chromosome 5: 18863848-18921839

Window Start End Accessions

4027 18863848 18894417 UKSE6_556 UKID57 Kil_0 UKSE6_278 UKID80 UKNW6_306 For_2 UKNW6_210

4028 18894447 18921839 UKSE6_556 UKNW6_210

==========

Haplotype 20211

Chromosome 5: 20201328-20251080

Window Start End Accessions

4083 20201328 20230958 Kil_0 Lc_0 Set_1 UKNW6_019 UKNW6_105 UKNW6_259 UKSE6_315 UKSW6_227

4084 20230981 20251080 Kil_0 UKNW6_259 UKSE6_315 Frd_1

==========

A3.3 HABITAT TYPE: GARDEN

Haplotype 223

Chromosome 1: 1701847-1742902

Window Start End Accessions

56 1701847 1720938 Edburgh_8 Sis_1 Igt_1 UKNW6_355 UKNW9_025 UKSW6_329 UKSE6_618 UKSW6_202 UKID87 UKID55 Hil_1 Cnt_1 UKNW9_010

57 1721173 1742902 Edburgh_8 Sis_1 Igt_1 UKNW6_355 UKNW9_025 UKSW6_329 UKSE6_618 UKSW6_202 UKID87 UKID55 Hil_1 Cnt_1 UKNW9_010

==========

Haplotype 1119

Chromosome 1: 7590135-7705277

Window Start End Accessions

263 7590135 7636992 02B6 09A3 Cnt_1 Edi_0 Ema_1A Ema_1B UKID103 UKID113 UKID33 UKID87 UKNW6_386 UKNW6_482 UKSE6_622 UKSE6_624 UKSE6_626 UKSE6_640 UKSW6_157 UKSW6_329

264 7637374 7657534 02B6 09A3 Cnt_1 Edi_0 Ema_1A Ema_1B UKID103 UKID113 UKID33 UKID87 UKNW6_386 UKNW6_482 UKSE6_622 UKSE6_624 UKSE6_626 UKSE6_640 UKSE6_192 UKID54 UKSW6_337 Edburgh_5 UKNW6_202 EM_183 Hil_1 Edi_1 EM_134 UKID65 Igt_1

265 7658135 7705277 02B6 09A3 Cnt_1 Edi_0 Ema_1A UKID103 UKID113 UKID33 UKID87 UKNW6_386 UKNW6_482 UKSE6_624 UKSE6_626 UKID65 UKID55 UKNW9_025 Sis_1 UKSE6_618 Edburgh_8

==========

Haplotype 3264

Chromosome 1: 20503582-20555114

Window Start End Accessions

670 20503582 20555114 UKID33 Igt_1 UKID87 UKID101 Cnt_1 UKID55 Sis_1 UKSW6_337 Ema_1A UKSE6_032

==========

Haplotype 3959

Chromosome 1: 23315350-23382634

Window Start End Accessions

787 23315350 23340472 09A3 Edi_1 Edburgh_5 EM_134 EM_183 UKID65 UKNW9_010 UKSE6_192 UKSE6_626

788 23341119 23359552 UKSE6_192 UKSE6_626 UKNW6_482

789 23359847 23382634 UKSE6_626 UKNW9_025 Edburgh_8 Igt_1

==========

Haplotype 4928

Chromosome 1: 27939409-27960918

Window Start End Accessions

976 27939409 27960918 Sis_1 Edburgh_8 UKID113 UKSE6_626 UKID64 Hil_1 UKSE6_624 UKSE6_618 09A3 UKID33 UKNW6_202 UKSW6_202 UKNW9_025

==========

Haplotype 6659

Chromosome 2: 8943655-9046384

Window Start End Accessions

1329 8943655 8988492 02B6 09A3 Cnt_1 Edburgh_8 Ema_1A Hil_1 Sis_1 UKID103 UKID113 UKID55 UKID87 UKID98 UKNW6_355 UKNW9_025 UKSE6_032 UKSE6_192 UKSE6_618 UKSE6_622 UKSE6_624 UKSE6_626 UKSE6_640 UKSW6_329

1330 8988713 9016106 02B6 09A3 Cnt_1 Edburgh_8 Ema_1A Hil_1 Sis_1 UKID103 UKID113 UKID55 UKID87 UKID98 UKNW6_355 UKNW9_025 UKSE6_032 UKSE6_192 UKSE6_618 UKSE6_622 UKSE6_624 UKSE6_626 UKSE6_640 UKSW6_329 UKSW6_157 EM_183

1331 9016562 9046384 02B6 Cnt_1 Edburgh_8 Ema_1A Hil_1 Sis_1 UKID103 UKID87 UKID98 UKNW9_025 UKSE6_618 UKSE6_622

==========

Haplotype 6941

Chromosome 2: 11544968-11596351

Window Start End Accessions

1401 11544968 11568561 UKNW6_355 UKNW6_202 UKSW6_337 UKNW9_010 Ema_1B UKSW6_329 UKID65 UKSW6_157 EM_183

1402 11569062 11596351 UKSW6_337 UKNW9_010 UKSE6_624

==========

Haplotype 6954

Chromosome 2: 11544968-11596351

Window Start End Accessions

1401 11544968 11568561 Edi_0 Edi_1 EM_134 Igt_1 UKID113 UKID33 UKID54 UKSE6_192

1402 11569062 11596351 Edi_0 Edi_1 EM_134 Igt_1 UKID113 UKID33 UKID54 UKSE6_192 EM_183

==========

Haplotype 6996

Chromosome 2: 11903599-12013466

Window Start End Accessions

1412 11903599 11954600 Edi_1 Edburgh_5 Ema_1B UKID54 UKID55 UKID65 UKID87 UKID98 UKSE6_626 UKSE6_640

1413 11954939 12013466 Edi_1 UKID54 UKSE6_626 UKSE6_640

==========

Haplotype 7505

Chromosome 2: 16569717-16665479

Window Start End Accessions

1529 16569717 16601757 Edburgh_5 Edburgh_8 Ema_1B Hil_1 Sis_1 UKID113 UKID98 UKNW6_355 UKNW9_010 UKNW9_025 UKSE6_192 UKSW6_157 UKSW6_329

1530 16603684 16639257 Hil_1 Sis_1 UKID113 UKID98 UKNW6_355 UKNW9_025 UKSW6_157 UKSW6_329 UKNW6_386 Ema_1A 02B6 UKSE6_618 EM_183 UKID87 UKID103

1531 16640924 16665479 Hil_1 Sis_1 UKID113 UKID98 UKNW6_355 UKNW9_025 UKSW6_157 UKSW6_329 UKNW6_386 Ema_1A 02B6 UKSE6_618 EM_183 UKID87 UKID103 UKID55 Cnt_1 UKNW6_482 UKID54 UKSE6_032 EM_134 UKID64

==========

Haplotype 8331

Chromosome 3: 3001989-3052696

Window Start End Accessions

1728 3001989 3018936 UKNW6_386 UKID113 UKSE6_192 UKID101 UKNW6_482 UKID98 UKSE6_622 EM_134 UKID64 Edi_0 UKID65 UKID54

1729 3018985 3052696 UKNW6_386 UKID113 UKSE6_192 UKID101 UKNW6_482 UKID98 UKSE6_622 EM_134 UKID64 Edi_0 UKID65 UKID54

==========

Haplotype 11733

Chromosome 3: 18383727-18484667

Window Start End Accessions

2328 18383727 18405985 02B6 Cnt_1 Edburgh_8 Ema_1A Hil_1 UKID101 UKID103 UKID55 UKID87 UKNW9_025 UKSE6_192 UKSE6_618 UKSE6_622 UKSE6_624 UKSE6_626

2329 18407136 18453900 02B6 Cnt_1 Edburgh_8 Ema_1A Hil_1 UKID103 UKID55 UKID87 UKNW9_025 UKSE6_618 UKSE6_622 UKSE6_624 UKSE6_626 Sis_1 UKNW6_355 EM_183 UKSE6_032 Igt_1 UKSW6_329

2330 18454342 18484667 02B6 Cnt_1 Edburgh_8 Ema_1A Hil_1 UKID103 UKID55 UKID87 UKNW9_025 UKSE6_618 UKSE6_622 UKSE6_624 UKSE6_626 Sis_1 EM_183 UKSE6_192 UKID101 Ema_1B EM_134 09A3

==========

Haplotype 12116

Chromosome 3: 20882525-20940708

Window Start End Accessions

2414 20882525 20912095 09A3 Ema_1B UKID113 UKID33 UKID54 UKID64 UKSE6_622 UKSW6_337

2415 20912771 20940708 09A3 UKID113 UKID33 UKID54 UKSE6_622 UKSW6_337 UKNW9_025 Sis_1 UKID101 Edi_1

==========

Haplotype 12113

Chromosome 3: 20837697-20940708

Window Start End Accessions

2413 20837697 20879578 EM_183

2414 20882525 20912095 EM_183 UKNW6_386 UKID55 UKID98 Sis_1 UKNW6_482 UKNW6_355 UKSE6_640 UKSW6_157 Edburgh_5 Ema_1A 02B6 UKSE6_626 UKSE6_618 UKNW9_010 UKID101 Hil_1 Edburgh_8 UKID65 UKSE6_624 UKID103 Igt_1 UKSW6_329

2415 20912771 20940708 UKID55 UKNW6_482 UKNW6_355 UKSE6_640 Edburgh_5 Ema_1A 02B6 UKSE6_626 UKSE6_618 UKNW9_010 Hil_1 Edburgh_8 UKID65 UKSE6_624 UKID103 Igt_1 UKSW6_329

==========

Haplotype 16953

Chromosome 5: 4574680-4632244

Window Start End Accessions

3416 4574680 4607113 Cnt_1 UKID55 UKSE6_640 UKSW6_337

3417 4607453 4632244 Cnt_1 UKID55 UKSE6_640 UKSW6_337 UKSE6_622 UKNW9_025 Sis_1 Ema_1A 02B6 UKSE6_618 EM_183 Hil_1 UKSE6_032 UKID65 UKID103

==========

Haplotype 17086

Chromosome 5: 5495017-5584934

Window Start End Accessions

3450 5495017 5531584 Edburgh_5

3451 5531783 5556750 UKSW6_157 Edburgh_5 Cnt_1 Sis_1 Edi_0 UKSE6_622 UKNW9_025 UKNW6_202 Hil_1 UKID55 UKSW6_202 Ema_1A UKSE6_618 UKID87 UKID33 02B6 UKNW9_010 UKID103

3452 5556965 5584934 UKSW6_157 Edburgh_5 Cnt_1 Sis_1 UKSE6_622 UKSE6_032 UKNW9_025 Hil_1 UKID55 UKSW6_202 Ema_1A UKSE6_618 UKID87 02B6 UKNW9_010 UKID103

==========

Haplotype 17354

Chromosome 5: 6911867-6953711

Window Start End Accessions

3512 6911867 6935750 09A3 Ema_1B Hil_1 UKID64 UKID65 UKID98 UKNW6_386 UKNW9_010 UKNW9_025 UKSE6_622 UKSE6_624 UKSE6_626 UKSW6_202

3513 6935945 6953711 09A3 Ema_1B Hil_1 UKID64 UKID65 UKID98 UKNW6_386 UKNW9_010 UKNW9_025 UKSE6_622 UKSE6_624 UKSE6_626 UKSW6_202 UKNW6_482 UKNW6_202 UKID33

==========

Haplotype 17357

Chromosome 5: 6911867-6953711

Window Start End Accessions

3512 6911867 6935750 UKID55 Edburgh_5 UKID101 UKID103 UKID54 UKSW6_337 02B6 UKID87 UKSE6_618 Sis_1 Ema_1A Cnt_1

3513 6935945 6953711 UKID55 Edburgh_5 UKID101 UKID103 UKID54 UKSW6_337 02B6 UKID87 UKSE6_618 Sis_1 Ema_1A Cnt_1

==========

Haplotype 19776

Chromosome 5: 18292316-18321487

Window Start End Accessions

3992 18292316 18307452 UKNW6_386 UKSE6_192 UKSW6_337 UKID54 UKID113 UKSW6_202 UKID101 Ema_1B 09A3

3993 18307733 18321487 UKSW6_202 UKID33

==========

Haplotype 21318

Chromosome 5: 26546023-26645553

Window Start End Accessions

4314 26546023 26566676 Edburgh_5 UKNW6_482 UKID65 UKSW6_337 UKID64 Cnt_1 Ema_1A 09A3 Igt_1 UKID87 UKSE6_624 UKSE6_618 UKID98

UKID103 Sis_1 Edi_0 UKNW6_355 UKSE6_622 UKNW9_025 UKNW9_010 UKID55 02B6

4315 26566823 26586649 Edburgh_5 UKNW6_482 UKID65 UKSW6_337 UKID64 Cnt_1 Ema_1A 09A3 Igt_1 UKID87 UKSE6_624 UKSE6_618 UKID98 UKID103 Sis_1 Edi_0 UKNW6_355 UKSE6_622 UKNW9_025 UKNW9_010 UKID55 02B6

4316 26586679 26606805 Edburgh_5 UKNW6_482 UKID65 UKSW6_337 UKSE6_640 UKSW6_329 UKID64 Cnt_1 Ema_1A 09A3 Igt_1 UKID87 UKSE6_624 UKSE6_618 UKID98 UKSW6_157 UKID103 Sis_1 Edi_0 UKNW6_355 UKID54 UKSE6_622 UKNW9_025 UKNW9_010 UKID55 02B6

4317 26607116 26628408 Edburgh_5 UKNW6_482 UKID65 UKSW6_337 UKSE6_640 UKSW6_329 UKID64 Cnt_1 Ema_1A 09A3 Igt_1 UKID87 UKSE6_624 Ema_1B UKSE6_618 UKID98 UKSW6_157 UKID103 Sis_1 Edi_0 UKNW6_355 UKID54 UKSE6_622 UKNW9_025 UKNW9_010 UKID55 02B6

4318 26628440 26645553 UKNW6_482 UKID65 UKSE6_640 UKSW6_329 UKID64 Cnt_1 Ema_1A 09A3 UKID87 UKSE6_624 Ema_1B UKSE6_618 UKID98 UKID103 Sis_1 Edi_0 UKID54 UKSE6_622 UKNW9_025 UKNW9_010 UKID55 02B6

==========

Haplotype 21361

Chromosome 5: 26766360-26864223

Window Start End Accessions

4324 26766360 26786498 UKSE6_640 UKID98 UKNW6_482 UKNW9_010 UKSW6_329 Ema_1B 02B6 UKSE6_192 UKID64 UKSE6_626 UKID103 UKNW9_025 UKID87 Ema_1A

4325 26786741 26819370 UKSE6_640 UKID98 UKSW6_202 UKNW6_482 UKNW9_010 UKSW6_329 Ema_1B 02B6 UKSE6_192 UKID64 UKSE6_626 UKID103 UKNW9_025 UKID87 Ema_1A

4326 26819749 26864223 UKSE6_640 UKID98 UKNW6_482 UKNW9_010 UKSW6_329 Ema_1B 02B6 UKSE6_192 UKID64 UKSE6_626 UKID103 UKNW9_025 UKID87 Ema_1A

==========

A3.4: HABITAT TYPE: OTHER

Haplotype 224

Chromosome 1: 1721173-1742902

Window Start End Accessions

57 1721173 1742902 CIBC_5B UKID28 CIBC_17 UKID109 Sq_1 UKID14 PHW_26 UKID39 CIBC_2 PHW_28 CIBC_5A HR_10 UKNW6_425 UKNW6_436 NFC_20

==========

Haplotype 714

Chromosome 1: 4902673-4970420

Window Start End Accessions

169 4902673 4940583 12A1 13B5 Alst_1 Asp_1 Boot_1 CIBC_5A Coc_1 Crl_1 HR_10 HR_5 Kent NFA_10 NFA_8 PHW_14 PHW_20 Poo_1 Sq_1 UKID109 UKID35 UKID39 UKID09 UKNW6_410 UKNW6_460 UKSE6_350 UKSE6_373 Ullapool3

170 4941411 4970420 12A1 13B5 Alst_1 CIBC_5A Crl_1 HR_5 NFA_10 NFA_8 Sq_1 UKID35 UKID09 UKNW6_410 UKSE6_350 UKSE6_373

==========

Haplotype 709

Chromosome 1: 4902673-4970420

Window Start End Accessions

169 4902673 4940583 PHW_13 UKID28 UKNW6_197 UKNW6_425 UKSW6_262 UKSW6_280 Ullapool4

170 4941411 4970420 PHW_13 UKNW6_197 UKSW6_262 UKSW6_280 Ullapool4 PHW_28 UKSW6_220 UKNW6_418 UKNW6_436 UKID48 UKID14

==========

Haplotype 903

Chromosome 1: 6043947-6109181

Window Start End Accessions

211 6043947 6077774 Kent 13B5 Sq_1 HR_5 UKID35 PHW_14 Coc_1 12A1 Boot_1 CIBC_5A Poo_1 Ullapool3 UKNW6_460 UKID109 UKSE6_350

212 6078621 6109181 Kent 13B5 Sq_1 HR_5 UKID35 PHW_14 Coc_1 12A1 Boot_1 CIBC_5A Poo_1 Ullapool3 UKNW6_460 UKID109 UKSE6_350

==========

Haplotype 3204

Chromosome 1: 19914749-19988033

Window Start End Accessions

657 19914749 19955458 Asp_1 Coc_1 NFC_20 Sq_8 UKID34 UKID09 UKNW6_460 UKSW6_280 Ullapool3

658 19955729 19988033 Coc_1 NFC_20 UKID34 UKNW6_460 UKSW6_280 Ullapool3 Poo_1 CIBC_2 UKSW6_220 UKNW6_425 UKSE6_351 Boot_1

==========

Haplotype 3203

Chromosome 1: 19914749-19988033

Window Start End Accessions

657 19914749 19955458 CIBC_17 CIBC_2 CIBC_4 HR_5 PHW_13 PHW_22 PHW_28 Sq_1 UKID35 UKID48 UKNW6_197 UKSE6_349

658 19955729 19988033 CIBC_17 CIBC_4 PHW_13 PHW_22 PHW_28 Sq_1 UKNW6_197 UKSE6_349 UKNW6_436 UKID39 UKID14

==========

Haplotype 3264

Chromosome 1: 20503582-20555114

Window Start End Accessions

670 20503582 20555114 UKID28 PHW_14 UKNW6_418 PHW_20 Ullapool4 UKSW6_070 UKID109 Sq_8 UKID48 12A1 HR_10 Crl_1 UKNW6_460 NFA_10 UKNW6_410

==========

Haplotype 4737

Chromosome 1: 26989512-27024099

Window Start End Accessions

943 26989512 27024099 Asp_1 UKNW6_460 Ullapool3 PHW_20 UKSE6_373 UKSE6_350 CIBC_4 UKNW6_197 UKID28 HR_5 UKID35 UKSE6_351 PHW_28 UKSW6_220 NFA_10 PHW_26 Poo_1

==========

Haplotype 4996

Chromosome 1: 28409994-28522644

Window Start End Accessions

988 28409994 28429071 13B5 CIBC_17 Crl_1 HR_5 Kent NFA_8 NFC_20 PHW_22 PHW_26 PHW_28 Sq_1 UKID35 UKSE6_349 UKSW6_070

989 28429158 28471823 Crl_1 Kent NFC_20 PHW_22 UKSW6_220 PHW_10

990 28472221 28522644 Crl_1 PHW_10 Alst_1 Poo_1 CIBC_5A PHW_13 UKNW6_436 UKNW6_425 PHW_20 UKID39 Sq_8 UKSE6_373 Boot_1 UKNW6_460

==========

Haplotype 6659

Chromosome 2: 8943655-9046384

Window Start End Accessions

1329 8943655 8988492 12A1 13B5 Asp_1 CIBC_2 CIBC_4 Crl_1 HR_10 HR_5 NFA_8 PHW_13 PHW_14 Sq_1 UKID35 UKNW6_197 UKNW6_410 UKSE6_350 UKSE6_351 UKSW6_070 Ullapool4

1330 8988713 9016106 12A1 13B5 Asp_1 CIBC_2 CIBC_4 Crl_1 HR_10 HR_5 NFA_8 PHW_13 PHW_14 Sq_1 UKID35 UKNW6_197 UKNW6_410 UKSE6_350 UKSE6_351 UKSW6_070 Ullapool4

1331 9016562 9046384 12A1 Crl_1 HR_10 HR_5 NFA_8 PHW_13 PHW_14 Sq_1 UKID35 UKNW6_410 UKSE6_350 UKSW6_070 Ullapool4

==========

Haplotype 6954

Chromosome 2: 11544968-11596351

Window Start End Accessions

1401 11544968 11568561 CIBC_17 CIBC_2 CIBC_4 CIBC_5B NFC_20 PHW_10 PHW_28 Sq_8 UKID28 UKID34 UKNW6_418 UKSE6_349 UKSE6_351

1402 11569062 11596351 CIBC_17 CIBC_2 CIBC_4 CIBC_5B NFC_20 PHW_10 PHW_28 Sq_8 UKID28 UKID34 UKNW6_418 UKSE6_349 UKSE6_351

==========

Haplotype 7725

Chromosome 2: 18083639-18143925

Window Start End Accessions

1580 18083639 18113697 12A1 13B5 Alst_1 Asp_1 Boot_1 CIBC_17 CIBC_2 CIBC_4 CIBC_5A CIBC_5B Coc_1 Crl_1 HR_10 HR_5 NFA_10 NFA_8 NFC_20 PHW_10 PHW_13 PHW_14 PHW_20 PHW_22 Poo_1 Sq_1 Sq_8 UKID109 UKID14 UKID28 UKID34 UKID35 UKID39 UKID48 UKID09 UKNW6_197 UKNW6_410 UKNW6_425 UKNW6_460 UKSE6_349 UKSE6_350 UKSE6_351 UKSE6_373 Ullapool3 Ullapool4

1581 18114045 18143925 12A1 13B5 Alst_1 Asp_1 Boot_1 CIBC_17 CIBC_2 CIBC_4 CIBC_5A CIBC_5B Coc_1 Crl_1 HR_10 HR_5 NFA_10 NFA_8 NFC_20 PHW_10 PHW_13 PHW_14 PHW_20 PHW_22 Poo_1 Sq_1 Sq_8 UKID109 UKID14 UKID28 UKID34 UKID35 UKID39 UKID48 UKID09 UKNW6_197 UKNW6_410 UKNW6_425 UKNW6_460 UKSE6_349 UKSE6_350 UKSE6_351 UKSE6_373 Ullapool3 Ullapool4 UKSW6_280 UKSW6_070

==========

Haplotype 8287

Chromosome 3: 2634191-2672381

Window Start End Accessions

1716 2634191 2653677 13B5 Boot_1 CIBC_4 CIBC_5A HR_5 NFA_10 PHW_22 Sq_1 UKID28 UKID34 UKID35 UKSE6_350 UKSW6_220

1717 2654177 2672381 13B5 Boot_1 CIBC_4 CIBC_5A HR_5 NFA_10 PHW_22 Sq_1 UKID28 UKID34 UKID35 UKSE6_350 UKSW6_220

==========

Haplotype 8284

Chromosome 3: 2634191-2672381

Window Start End Accessions

1716 2634191 2653677 12A1 Alst_1 Asp_1 CIBC_2 Coc_1 HR_10 PHW_13 PHW_14 PHW_20 PHW_26 PHW_28 Poo_1 Sq_8 UKID14 UKNW6_197 UKNW6_410 UKNW6_425 Ullapool3 Ullapool4

1717 2654177 2672381 12A1 Alst_1 Asp_1 CIBC_2 HR_10 PHW_13 PHW_14 PHW_26 PHW_28 Sq_8 UKNW6_197 UKNW6_410 Ullapool3 Ullapool4 UKSW6_070 UKSE6_373

==========

Haplotype 8587

Chromosome 3: 4296691-4382190

Window Start End Accessions

1774 4296691 4324868 UKID109 Ullapool3 UKID34 Crl_1 UKSE6_351 CIBC_17 UKSE6_349 NFA_8 UKID28

1775 4325283 4349319 Ullapool3 UKID34 UKID09 NFA_10 Crl_1 UKSW6_280 CIBC_5A 13B5 UKSE6_351 CIBC_17 UKSE6_349 NFA_8 UKID28 UKSW6_262 Sq_1 Coc_1 PHW_20

1776 4350669 4382190 UKNW6_460 Ullapool3 UKID34 Crl_1 UKSW6_280 UKSE6_351 CIBC_17 UKSE6_349 NFA_8 UKID28 UKSW6_262 PHW_20

==========

Haplotype 8658

Chromosome 3: 4940454-5068238

Window Start End Accessions

1795 4940454 4967281 Sq_1 NFA_10 Coc_1 UKID48 HR_5 PHW_20 13B5 12A1 Alst_1 PHW_14 UKNW6_436 UKID09 CIBC_5A UKSE6_350

1796 4967935 4993799 Sq_1 UKID109 NFA_10 Coc_1 UKID48 HR_5 PHW_20 13B5 12A1 Kent UKID34 Alst_1 PHW_14 UKNW6_436 UKID09 CIBC_5A UKSE6_350

1797 4995760 5032864 Sq_1 UKID109 NFA_10 Coc_1 UKID48 HR_5 PHW_20 13B5 12A1 Kent UKID34 Alst_1 UKSW6_220 PHW_14 UKNW6_436 UKID09 CIBC_5A UKSE6_350

1798 5035207 5068238 Sq_1 NFA_10 Coc_1 UKID48 HR_5 PHW_20 13B5 12A1 UKID34 Alst_1 PHW_14 UKNW6_436 UKID09 CIBC_5A UKSE6_350

==========

Haplotype 9040

Chromosome 3: 7185432-7236180

Window Start End Accessions

1883 7185432 7205233 12A1 CIBC_5A PHW_13 PHW_22 UKNW6_197 UKNW6_425 UKSE6_350 UKSW6_070 UKSW6_220

1884 7206226 7236180 CIBC_5A PHW_22 UKID35 UKID09 UKNW6_418 UKID48

==========

Haplotype 11733

Chromosome 3: 18383727-18484667

Window Start End Accessions

2328 18383727 18405985 12A1 13B5 Asp_1 CIBC_5A Crl_1 HR_10 NFA_10 NFA_8 NFC_20 PHW_14 UKID09 UKNW6_410 UKNW6_436 UKSE6_373 UKSW6_070

2329 18407136 18453900 12A1 13B5 Asp_1 CIBC_5A Crl_1 HR_10 NFA_10 NFA_8 PHW_14 UKID09 UKNW6_410 UKNW6_436 UKSE6_373 UKSW6_070 PHW_13 UKNW6_197 Ullapool4

2330 18454342 18484667 12A1 13B5 Asp_1 CIBC_5A Crl_1 HR_10 NFA_10 NFA_8 PHW_14 UKID09 UKNW6_410 UKNW6_436 UKSE6_373 UKSW6_070 PHW_28 Coc_1 UKID39 CIBC_17 PHW_10

==========

Haplotype 12113

Chromosome 3: 20882525-20940708

Window Start End Accessions

2414 20882525 20912095 NFA_10 UKID35 CIBC_4 PHW_28 Alst_1 Poo_1 UKID09 PHW_13 UKNW6_436 UKNW6_425 UKSE6_350 CIBC_5B UKNW6_197 NFC_20 Sq_8 Boot_1 UKID48 PHW_14 HR_5 UKID34 Ullapool4 UKID109 Ullapool3 UKID14 HR_10

2415 20912771 20940708 NFA_10 UKID35 CIBC_4 Alst_1 Poo_1 UKID09 UKNW6_436 UKNW6_425 CIBC_5B UKNW6_197 NFC_20 Boot_1 UKID48 PHW_14 Ullapool4 UKID109 Ullapool3 UKID14 HR_10

==========

Haplotype 12758

Chromosome 4: 1064420-1123843

Window Start End Accessions

2555 1064420 1080424 12A1 13B5 Alst_1 Asp_1 Boot_1 CIBC_2 CIBC_4 CIBC_5A CIBC_5B Coc_1 Crl_1 HR_10 HR_5 Kent NFA_10 NFA_8 PHW_13 PHW_14 PHW_22 PHW_26 PHW_28 Poo_1 Sq_1 UKID109 UKID14 UKID28 UKID35 UKID39 UKID48 UKID09 UKNW6_197 UKNW6_410 UKNW6_418 UKNW6_425 UKNW6_436 UKNW6_460 UKSE6_349 UKSE6_350 UKSE6_373 UKSW6_070 UKSW6_220 Ullapool3 Ullapool4

2556 1081604 1112186 Alst_1 Asp_1 CIBC_2 CIBC_4 CIBC_5B Crl_1 HR_10 HR_5 NFA_10 NFA_8 PHW_13 PHW_14 UKID14 UKID35 UKID39 UKID48 UKID09 UKSE6_349 UKSE6_373 UKSW6_070 UKSE6_351 PHW_10

2557 1112750 1123843 Asp_1 CIBC_2 Crl_1 HR_10 NFA_10 NFA_8 PHW_13 PHW_14 UKID14 UKID35 UKID39 UKID48 UKID09 UKSE6_349 UKSE6_373 UKSW6_220 UKNW6_436

==========

Haplotype 16336

Chromosome 5: 465136-477734

Window Start End Accessions

3272 465136 477734 UKSW6_220 UKNW6_425 UKSW6_280 Poo_1 UKID09 Coc_1 Boot_1 UKID39 Alst_1 PHW_20 UKID14 Sq_8 UKNW6_436 PHW_26 UKNW6_460

==========

Haplotype 16953

Chromosome 5: 4574680-4632244

Window Start End Accessions

3416 4574680 4607113 13B5 Crl_1 HR_10 NFA_8 UKID35 UKID48 UKNW6_410 UKSW6_070

3417 4607453 4632244 13B5 Crl_1 HR_10 NFA_8 UKID35 UKID48 UKNW6_410 UKSW6_070 UKID09 Coc_1 UKID28 UKSE6_350 UKSE6_373 PHW_14

==========

Haplotype 17099

Chromosome 5: 5531783-5584934

Window Start End Accessions

3451 5531783 5556750 13B5 CIBC_5A Kent NFA_10 PHW_26 Sq_1 UKID28 UKID34 UKSW6_220 UKSW6_262 UKSW6_280

3452 5556965 5584934 Kent PHW_26 UKID28 UKID34 UKSW6_220 UKSW6_262 UKSW6_280

==========

Haplotype 17093

Chromosome 5: 5531783-5584934

Window Start End Accessions

3451 5531783 5556750 UKNW6_197 NFC_20 PHW_13 Alst_1 PHW_10 UKNW6_436 Ullapool4 UKSE6_351 Ullapool3 PHW_20 UKNW6_460 UKSE6_373

3452 5556965 5584934 UKNW6_197 NFC_20 PHW_13 Alst_1 PHW_10 UKNW6_436 Ullapool4 UKSE6_351 Ullapool3 PHW_20 UKNW6_460 UKSE6_373

==========

Haplotype 17086

Chromosome 5: 5531783-5584934

Window Start End Accessions

3451 5531783 5556750 NFA_8 UKID09 UKSW6_070 12A1 Poo_1 HR_10 PHW_14 Coc_1 UKNW6_410 UKSE6_350 UKID35 HR_5 Crl_1 UKID14 UKID39

3452 5556965 5584934 NFA_8 UKID09 UKSW6_070 12A1 Poo_1 HR_10 PHW_14 Coc_1 UKNW6_410 UKID35 HR_5 CIBC_5B Crl_1 UKID14 UKID39

==========

Haplotype 17357

Chromosome 5: 6911867-6953711

Window Start End Accessions

3512 6911867 6935750 UKNW6_410 Alst_1 13B5 Sq_1 UKSE6_350 PHW_20 NFA_10 PHW_14 Crl_1 CIBC_5A HR_10 NFA_8 Coc_1

3513 6935945 6953711 UKNW6_410 Alst_1 13B5 Sq_1 PHW_20 NFA_10 PHW_14 Crl_1 CIBC_5A HR_10 NFA_8 Coc_1 12A1 UKSW6_070

==========

Haplotype 17355

Chromosome 5: 6911867-6953711

Window Start End Accessions

3512 6911867 6935750 CIBC_17 CIBC_2 CIBC_4 CIBC_5B Kent NFC_20 PHW_13 PHW_28 UKID109 UKID28 UKID34 UKNW6_197 UKNW6_418 Ullapool4

3513 6935945 6953711 CIBC_17 CIBC_2 CIBC_4 CIBC_5B Kent NFC_20 PHW_13 PHW_28 UKID109 UKID28 UKID34 UKNW6_197 UKNW6_418 Ullapool4 Asp_1

==========

Haplotype 17354

Chromosome 5: 6911867-6953711

Window Start End Accessions

3512 6911867 6935750 12A1 Boot_1 HR_5 PHW_22 Poo_1 UKID14 UKID35 UKID39 UKID48 UKID09 UKNW6_425 UKNW6_460 UKSE6_349 UKSW6_220

3513 6935945 6953711 Boot_1 HR_5 PHW_22 Poo_1 UKID14 UKID35 UKID39 UKID48 UKID09 UKNW6_425 UKNW6_460 UKSE6_349 UKSW6_220 UKNW6_436 UKSE6_351 UKSE6_373 Ullapool3

==========

Haplotype 19926

Chromosome 5: 18740583-18779109

Window Start End Accessions

4021 18740583 18759801 12A1 13B5 CIBC_5A Coc_1 Crl_1 HR_10 HR_5 NFA_10 NFA_8 PHW_13 PHW_14 PHW_22 Sq_1 UKID35 UKID48 UKNW6_197 UKNW6_460 UKSE6_350 UKSW6_070

4022 18760126 18779109 12A1 13B5 CIBC_5A Coc_1 Crl_1 HR_10 HR_5 NFA_10 NFA_8 PHW_13 PHW_14 PHW_22 Sq_1 UKID35 UKID48 UKNW6_197 UKNW6_460 UKSE6_350 UKSW6_070 UKSW6_220

==========

Haplotype 20165

Chromosome 5: 19900567-20007982

Window Start End Accessions

4071 19900567 19934412 Sq_1 PHW_14 UKSE6_350 UKID35 HR_10 HR_5 UKSE6_373 PHW_22 12A1 UKID39 NFA_10 UKSW6_280 UKNW6_460 NFC_20 Ullapool4 UKNW6_410 Boot_1 Coc_1

4072 19934670 19961212 UKID09 NFC_20 UKSW6_220 PHW_26 UKNW6_418 Ullapool4 UKNW6_410 Boot_1 Coc_1

4073 19961430 19982849

4074 19983056 20007982 UKID09

==========

Haplotype 20162

Chromosome 5: 19900567-19934412

Window Start End Accessions

4071 19900567 19934412 CIBC_17 UKID28 UKSW6_220 UKNW6_418 UKNW6_197 PHW_13 PHW_20 PHW_28 CIBC_5A Kent PHW_26 Sq_8 CIBC_5B UKID09

==========

Haplotype 21277

Chromosome 5: 26374287-26416718

Window Start End Accessions

4305 26374287 26393222 PHW_26 UKSE6_373 UKNW6_460 NFA_10 Coc_1 HR_10 Asp_1 Boot_1 12A1 UKSE6_350 Sq_1 UKNW6_425 UKID34

4306 26394590 26416718 PHW_26 UKSE6_373 UKNW6_460 NFA_10 Coc_1 HR_10 Asp_1 Boot_1 12A1 UKSE6_350 Sq_1 UKNW6_425 UKID34

==========

Haplotype 21291

Chromosome 5: 26374287-26416718

Window Start End Accessions

4305 26374287 26393222 PHW_28 UKID39

4306 26394590 26416718 CIBC_4 Alst_1 CIBC_5A UKSE6_351 CIBC_5B Sq_8 UKID48 PHW_14 13B5 PHW_10

==========

Haplotype 21318

Chromosome 5: 26546023-26645553

Window Start End Accessions

4314 26546023 26566676 UKID14 Ullapool4 PHW_13 UKNW6_425 UKID48 UKID109 PHW_22 Boot_1 Ullapool3 Poo_1 HR_5 PHW_10 UKSE6_350 UKNW6_197 Alst_1 UKNW6_418 UKID35 UKID28 UKNW6_410 Crl_1 UKSE6_351 NFA_8

4315 26566823 26586649 UKID14 Ullapool4 PHW_13 UKNW6_425 UKID48 UKID109 PHW_22 Boot_1 Ullapool3 Poo_1 HR_5 PHW_10 UKSE6_350

UKNW6_197 Alst_1 UKNW6_418 UKID35 UKID28 UKNW6_410 Crl_1 UKSE6_351 NFA_8

4316 26586679 26606805 UKID14 Ullapool4 12A1 UKSE6_373 PHW_13 UKID34 UKNW6_425 UKID48 UKID109 PHW_22 Boot_1 PHW_20 NFA_10 Ullapool3 Sq_1 Poo_1 HR_5 PHW_10 UKSE6_350 UKNW6_197 Alst_1 UKNW6_418 UKID35 UKID28 HR_10 UKNW6_410 Crl_1 UKSE6_351 NFA_8

4317 26607116 26628408 UKID14 Ullapool4 UKSE6_373 PHW_13 UKID34 UKNW6_425 UKID48 UKID109 PHW_22 Boot_1 PHW_20 NFA_10 Ullapool3 Sq_1 Poo_1 HR_5 PHW_10 UKSE6_350 UKNW6_197 Alst_1 UKNW6_418 PHW_26 UKID35 UKID28 HR_10 UKNW6_410 Crl_1 UKSE6_351 NFA_8

4318 26628440 26645553 UKID14 UKSE6_373 PHW_13 UKID34 UKNW6_425 UKID48 UKID109 PHW_22 Boot_1 PHW_20 NFA_10 Sq_1 Poo_1 HR_5 UKSE6_350 Alst_1 UKNW6_418 PHW_26 UKID35 HR_10 UKNW6_410 12A1 Crl_1 NFA_8

==========

A4: DISTRIBUTION OF GENOTYPE CLUSTERS ACROSS HABITAT TYPES

Genotype Cluster	Accession Name	Lat.	Long.	Location	Wall/ outcrop	Garden	Railway	Other
UK-Scandanavian	UKID120	56.7	-6	Ardtoe, Scotland	x			
	Ty0	56.4	-5.2	Taynuilt, Scotland	x			
	Mc0	54.6	-2.3	Mickells Fell, Pennines	x			
	UKNW6170	54.6	-3.1	Brathay to Hawkshead, Cumbria	x			
	UKNW6306	54.6	-3.1	Borrowdale, Cumbria	x			
	UKNW6178	54.4	-3	Brathay to Hawkshead, Cumbria	x			
	UKNW6202	54.4	-3	Brathay village, Cumbria		x		
	Bur-0	54.1	-6.2	Burren, Ireland	x			
	UKSE6565	51.3	1.1	Canterbury, Kent	x			
	UKSE6581	51.3	1.1	Fordwich, Kent	x			
UK-French	Ullapool3	57.9	-5.2	Ullapool, Scotland				x
	UKID57	56.5	0.4	Peebles, Scotland	x			
	Edinburgh5	56	-3.2	Edinburgh, Scotland		x		
	09A3	56	-3.2	Edinburgh, Scotland		x		
	UKID9	55.6	-3.5	Biggar, Scotland				x
	UKID39	55.4	-2.8	Hawick, Scotland				x
	UKID65	55.3	-1.9	Rothbury, Northumberland		x		
	UKID14	55.2	-2	Cambo, Northumberland				x
	Asp1	54.8	-3.3	Aspatria, Cumbria				x
	UKID48	54.7	-2.7	Lazonby, Cumbria				x
	Coc1	54.7	-3.4	Cockermouth, Cumbria				x
	UKNW6425	54.7	-3.4	Cockermouth, Cumbria				x
	UKNW6436	54.7	-3.4	Cockermouth, Cumbria				x
	UKNW6460	54.7	-3.4	Cockermouth, Cumbria				x
	Poo1	54.6	-2.8	Pooley Bridge, Cumbria				x
	UKNW6355	54.6	-3.1	Keswick, Cumbria		x		
	UKNW6386	54.6	-3.1	Keswick, Cumbria		x		
	UKNW9010	54.6	-3.1	Keswick, Cumbria		x		
	UKNW6259	54.6	-3.3	Scawgill Bridge, Cumbria	x			
	UKNW6482	54.4	-2.9	Windemere, Cumbria		x		
	UKNW6050	54.4	-3	Ambleside to Rydal rd, Cumbria	x			
	UKNW6079	54.4	-3	Ambleside to Rydal rd, Cumbria	x			
	UKNW6101	54.4	-3	Outgate, Cumbria	x			
	UKNW6105	54.4	-3	Outgate, Cumbria	x			
	Boot1	54.4	-3.3	N.A.				x
	UKID98	52.3	-1.6	Kenilworth, Warwickshire		x		
	UKID28	52.3	-1.7	Dunich Heath, Suffolk				x
	UKID108	52.1	-2.3	Malvern Hill, Worcestershire	x			
	UKID109	52	-2.4	Ledbury, Herefordshire				x
	PHW31	51.5	-3.2	Ely, Cambridgeshire			x	
	PHW22	51.4	-1.7	Marlborough, Wiltshire				x
	CIBC5	51.4	-0.6	N.A.				x
	NFA10	51.4	-0.6	N.A.				x
	UKID64	51.3	1	Rough Common, Kent		x		
	EM183	51.3	0.5	East Malling Research, Kent		x		
	UKSE6032	51.3	0.5	East Malling Research, Kent		x		
	UKSE6373	51.3	0.4	Wateringbury, Kent				x
	Igt1	51.3	0.3	Igtham, Kent		x		

Genotype Cluster	Accession Name	Lat.	Long.	Location	Wall/ outcrop	Garden	Railway	Other
	UKSE6624	51.1	0.4	Scotney Castle, East Sussex		x		
	UKSE6626	51.1	0.4	Scotney Castle, East Sussex		x		
	UKSE6640	51.1	0.4	Scotney Castle, East Sussex		x		
	UKSW6070	50.4	-4.7	Cornwall				x
	UKSW6157	50.4	-4.7	Luxulyan, Cornwall		x		
	UKSE6624	51.1	0.4	Scotney Castle, East Sussex		x		
	UKSE6626	51.1	0.4	Scotney Castle, East Sussex		x		
	UKSE6640	51.1	0.4	Scotney Castle, East Sussex		x		
	UKSW6157	50.4	-4.7	Luxulyan, Cornwall		x		
	UKSW6227	50.4	-4.9	St Dennis, Cornwall	x			
	UKSW6337	50.3	-4.6	Fowey, Cornwall		x		
UK-Iberian/French	Lc0	57	-4	Loch Carron, Scotland	x			
	For2	56.6	-4.1	Fortingall, Scotland	x			
	Edi0	56	-3	Edinburgh, Scotland		x		
	Kil0	55.6	-5.7	Killean, Scotland	x			
	UKID80	54.7	-2.9	Unthank, Cumbria	x			
	UKNW6418	54.7	-3.4	Cockermouth, Cumbria				x
	UKNW6019	54.4	-3	Ambleside to Rydal rd, Cumbria	x			
	UKNW6040	54.4	-3	Ambleside to Rydal rd, Cumbria	x			
	UKID33	54.1	-1.6	Fountains Abbey, Lincolnshire		x		
	Set1	54.1	-2.3	Settle, Pennines	x			
	UKID101	53.2	-1.4	Hardwick Hall		x		
	UKID54	52.6	1.3	Norwich, Norfolk		x		
	UKID113	52.3	-1.5	Warwick castle, Warwickshire	x			
	UKID34	52.2	1.4	Farnham, Suffolk				x
	UKSE6315	52.2	-1.7	West Malling village, Kent	x			
	UKID58	51.6		Paddock Wood, Kent			x	
	CIBC52	51.4	-0.6	CIBC				x
	CIBC17	51.4	-0.6	CIBC				x
	CIBC2	51.4	-0.6	CIBC				x
	CIBC4	51.4	-0.6	CIBC				x
	NFC20	51.4	-0.6	NFC				x
	Sq8	51.4	-0.6	SQ				x
	Frd1	51.3	1.1	Fordwich, Kent	x			
	UKSE6597	51.3	1.1	Fordwich, Kent	x			
	EM134	51.3	0.5	East Malling Research, Kent		x		
	UKSE6192	51.3	0.5	East Malling Research, Kent		x		
	UKSE6272	51.3	0.4	East Malling village, Kent	x			
	UKSE6278	51.3	0.4	East Malling village, Kent	x			
	UKSE6349	51.3	0.4	Wateringbury, Kent				x
	UKSE6351	51.3	0.4	Wateringbury, Kent				x
	PHW10	51.3	0.1	Kent (PHW)				x
	UKID15	51.2	1	Chilham (BR Station), Kent			x	
	Kent	51.2	0.4	Kent				x
	PHW26	50.7	-3.8	Chagford				x
	UKSW6202	50.4	-4.9	St Columb, Cornwall		x		
	PHW28	50.4	-3.6	Dartmode, Devon				x
UK-German	Kyl1	57.3	-5.7	Kyle of Localsh, Scotland	x			
	Edi1	55.9	-3.2	Edinburgh, Scotland		x		
	UKSE6254	51.3	0.5	East Malling village, Kent	x			
	UKSW6262	50.3	-4.9	St Stephens, Cornwall				x
	UKSW6280	50.3	-4.9	St Stephens, Cornwall				x

Genotype Cluster	Accession Name	Lat.	Long.	Location	Wall/ outcrop	Garden	Railway	Other
UK only	Ullapool4	57.9	-5.2	Ullapool, Scotland				x
	Edinburgh8	56	-3.2	Edinburgh, Scotland		x		
	02B6	56	-3.2	Edinburgh, Scotland		x		
	Crl1	54.9	-2.9	Carlisle, Cumbria				x
	Alst1	54.8	-2.4	Alst				x
	UKNW6410	54.7	-3.4	Cockermouth, Cumbria				x
	UKNW9025	54.6	-3.1	Keswick, Cumbria		x		
	UKNW6078	54.4	-3	Ambleside to Rydal rd, Cumbria	x			
	UKNW6210	54.4	-3	Grasmere, Cumbria	x			
	UKNW6197	54.4	-3	Skelfold rd, Cumbria				x
	UKID55	53	-1.1	Nottingham		x		
	UKID103	51.8	-0.5	Whipsnade Zoo, Bucks		x		
	HR5	51.4	-0.6	HR				x
	NFA8	51.4	-0.6	NFA				x
	Sq1	51.4	-0.6	SQ				x
	UKID17	51.4	0.1	Chiselhurst, Kent			x	
	UKSE6544	51.3	1.1	Canterbury, Kent	x			
	UKSE6556	51.3	1.1	Canterbury, Kent	x			
	Cnt1	51.3	1.1	Canterbury, Kent		x		
	UKID35	51.3	0.9	Faversham, Kent				x
	Ema1	51.3	0.5	East Malling Research, Kent		x		
	UKSE6350	51.3	0.4	Wateringbury, Kent				x
	Wis1	51.3	-0.5	Wisley Garden		x		
	PHW13	51.3	0.1	Kent (PHW)				x
	PHW14	51.3	0.1	Kent (PHW)				x
	Sis1	51.1	0.6	Sissinghurst garden, Kent		x		
	UKSE6618	51.1	0.4	Scotney Castle, East Sussex		x		
	UKSE6622	51.1	0.4	Scotney Castle, East Sussex		x		
	Hil1	51	-1.5	Hillier Arboretum		x		
	UKID87	50.8	-0.7	Bognor Regis		x		
	UKSW6025	50.4	-4.7	Twydreath, Cornwall	x			
	UKSW6329	50.3	-4.8	St Austel, Cornwall		x		

A5: PCA AND STRUCTURE GENOTYPE CLUSTERS

A5.1 PCA GENOTYPE CLUSTERS

UK-Scandinavian	UK-German	UK-US-Iberian-French	UK-French	UK-only
Bur_0	Edi_1	CIBC_17	09A3	02B6
Mc_0	Kyl_1	CIBC_2	Asp_1	12A1
Ty_0	UKSE6_254	CIBC_4	Boot_1	13B5
UKID120	UKSW6_262	CIBC_5B	CIBC_5A	Alst_1
UKNW6_170	UKSW6_280	Edi_0	Coc_1	Cnt_1
UKNW6_178		EM_134	Edburgh_5	Crl_1
UKNW6_202		For_2	EM_183	Edburgh_8
UKNW6_306		Frd_1	Ema_1B	Ema_1A
UKSE6_565		Kent	Igt_1	Hil_1
UKSE6_581		Kil_0	NFA_10	HR_10
		Lc_0	PHW_20	HR_5
		NFC_20	PHW_22	NFA_8
		PHW_10	PHW_31	PHW_13
		PHW_26	Poo_1	PHW_14
		PHW_28	UKID09	Sis_1
		Set_1	UKID108	Sq_1
		Sq_8	UKID109	UKID103
		UKID101	UKID14	UKID17
		UKID113	UKID28	UKID35
		UKID15	UKID39	UKID55
		UKID33	UKID48	UKID87
		UKID34	UKID57	UKNW6_078
		UKID54	UKID64	UKNW6_197
		UKID58	UKID65	UKNW6_210
		UKID72	UKID98	UKNW6_410
		UKID80	UKNW6_050	UKNW9_025
		UKNW6_019	UKNW6_079	UKSE6_350
		UKNW6_040	UKNW6_101	UKSE6_544
		UKNW6_418	UKNW6_105	UKSE6_556
		UKSE6_192	UKNW6_259	UKSE6_618
		UKSE6_272	UKNW6_355	UKSE6_622
		UKSE6_278	UKNW6_386	UKSW6_025
		UKSE6_315	UKNW6_425	UKSW6_329
		UKSE6_349	UKNW6_436	Ullapool4
		UKSE6_351	UKNW6_460	Wis_1
		UKSE6_597	UKNW6_482	
		UKSW6_202	UKNW9_010	
		UKSW6_220	UKSE6_032	
			UKSE6_373	
			UKSE6_624	
			UKSE6_626	

	UKSE6_640
	UKSW6_070
	UKSW6_157
	UKSW6_227
	UKSW6_337
	Ullapool3

A5.2 STRUCTURE GENOTYPE CLUSTERS

Real UK population - k=5				
Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Mc_0	12A1	Boot_1	Asp_1	CIBC_2
Ty_0	13B5	Edi_1	CIBC_5A	CIBC_4
UKID120	Alst_1	Ema_1B	Coc_1	CIBC_5B
UKNW6_170	Cnt_1	Kent	Edburgh_5	Edi_0
UKNW6_178	Crl_1	Kyl_1	PHW_20	EM_134
UKNW6_202	Edburgh_8	Lc_0	PHW_22	For_2
UKNW6_306	EM_183	Poo_1	PHW_31	Frd_1
UKSE6_565	Ema_1A	Set_1	UKID108	Igt_1
UKSE6_581	Hil_1	UKID14	UKID109	Kil_0
	HR_10	UKNW6_425	UKID39	NFC_20
	HR_5	UKNW9_010	UKID48	PHW_10
	NFA_10		UKID57	PHW_26
	NFA_8		UKID64	PHW_28
	PHW_13		UKID65	Sq_8
	PHW_14		UKNW6_050	UKID101
	Sis_1		UKNW6_079	UKID113
	Sq_1		UKNW6_101	UKID15
	UKID09		UKNW6_105	UKID28
	UKID103		UKNW6_259	UKID33
	UKID17		UKNW6_386	UKID34
	UKID35		UKNW6_436	UKID54
	UKID55		UKNW6_460	UKID58
	UKID87		UKNW6_482	UKID72
	UKID98		UKSE6_373	UKID80
	UKNW6_078		UKSE6_624	UKNW6_019
	UKNW6_197		UKSE6_626	UKNW6_040
	UKNW6_210		UKSE6_640	UKNW6_418
	UKNW6_355		UKSW6_202	UKSE6_192
	UKNW6_410		UKSW6_227	UKSE6_254
	UKNW9_025		UKSW6_337	UKSE6_272
	UKSE6_032		Ullapool3	UKSE6_278
	UKSE6_350			UKSE6_311
	UKSE6_544			UKSE6_315
	UKSE6_556			UKSE6_349
	UKSE6_618			UKSE6_351
	UKSE6_622			UKSE6_597
	UKSW6_025			UKSW6_220
	UKSW6_070			UKSW6_262
	UKSW6_157			UKSW6_280
	UKSW6_329			
	Ullapool4			
	Wis_1			

A6: GENES POSSESSING SIGNATURES OF SELECTION

A6.1: NB-LRR, RLK AND RLP GENES

Gene	Type	Wall	Garden	Other	All UK	Ka (non-syn)	Ks (syn)	Ka/Ks ratio	Crossover rate
AT1G01780	PF00412: LIM domain	*				0	1	0:1	0.76799
AT1G05700	kinase-TM-LRR		*			2	5	2:5	0.59704
AT1G05760				*		0	3	0:1	1.1041
AT1G14370	cytoplasmic RLK			*		1	0	1:0	0.71537
AT1G14390	kinase-TM-LRR			*		5	5	1:1	0.32711
AT1G17600	NBS-LRR			*		11	5	11:5	0.01871
AT1G17610	TIR-NBS			*		2	1	2:1	0.01601
AT1G21750	CXHC_Ath08	*	*			7	8	7:8	0.41942
AT1G53420	kinase-TM-LRR			*		23	18	23:18	0.48997
AT1G53430	kinase-TM-LRR			*		9	23	9:23	0.4379
AT1G53440	kinase-TM-LRR			*		6	5	6:5	0.16195
AT1G53730	kinase-TM-LRR					3	18	1:6	7.58143
AT1G55020	lipoxygenase		*	*	*	1	4	1:4	0.53129
AT1G62950	kinase-TM-LRR		*			9	9	1:1	3.27546
AT1G71830	kinase-TM-LRR			*		0	41	0:1	0.87785
AT1G72300	kinase-TM-LRR	*				12	8	3:2	1.06269
AT1G74360	kinase-TM-LRR		*			4	4	1:1	3.34934
AT1G75820	kinase-TM-LRR			*		9	23	9:23	1.17651
AT2G20850	kinase-TM-LRR		*	*		5	4	5:4	0.28439
AT2G26290	cytoplasmic RLK	*				4	7	4:7	3.25695
AT2G27060	kinase-TM-LRR		*	*		20	14	10:7	0.92419
AT2G28010	CXHC_Ath02		*			6	1	6:1	0.06562
AT2G28040	CXHC_Ath01		*			1	4	1:4	0.17294
AT2G39380	PF00412: LIM domain		*			1	0	1:0	0.856
AT2G43690	L-Lectin			*		5	3	5:3	0.69647
AT3G04370	CHXC_Ath02	*				7	4	7:4	0.24413
AT3G08680	kinase-TM-LRR			*		2	6	1:3	0.31789
AT3G09830	cytoplasmic RLK		*			4	4	1:1	3.21973
AT3G13380	kinase-TM-LRR			*		4	7	4:7	0.39926
AT3G14460	NBS-LRR				*	20	25	4:5	0.37005
AT3G14470	NBS-LRR				*	28	28	1:1	0.37033
AT3G14840	kinase-TM-LRR			*		0	0	0	0.4679
AT3G20600				*		3	2	3:2	0.33972
AT3G49670	kinase-TM-LRR	*	*	*	*	5	4	5:4	0.28812

AT3G49750	TM-LRR	*	*	*	*	0	2	0:1	2.14613
AT3G56370	kinase-TM-LRR		*	*	*	8	10	4:5	1.39897
AT4G02420	L-Lectin			*		7	27	7:27	1.28406
AT4G17780	F-box myb transcription factor	*				10	0	1:0	0.43791
AT4G19530	NBS-LRR	*			*	18	8	9:4	0.68985
AT4G20270	kinase-TM-LRR	*				10	3	10:3	3.74879
AT4G20380					*	0	2	0:1	0.34255
AT5G02290	cytoplasmic RLK			*		3	6	1:2	0.4171
AT5G14210	kinase-TM-LRR	*	*	*	*	6	17	6:17	0.62926
AT5G15080	cytoplasmic RLK	*				1	2	1:2	0.19834
AT5G16000	kinase-TM-LRR	*				2	2	1:1	6.69866
AT5G16900	kinase-TM-LRR		*	*		25	16	25:16	0.27451
AT5G20480	kinase-TM-LRR	*	*	*	*	42	17	42:17	0.54442
AT5G45230	NBS-LRR		*			8	6	4:3	0.30901
AT5G45770	TM-LRR				*	11	7	11:7	1.79169
AT5G46260	NBS-LRR	*		*		53	34	53:34	1.69334
AT5G46510	NBS-LRR	*				13	3	13:3	0.30363
AT5G46520	NBS-LRR	*				34	10	17:5	0.29781
AT5G49140	NBS-LRR			*		22	1	22:1	1.25729
AT5G49760	kinase-TM-LRR	*				7	21	1:3	3.93117
AT5G49770	kinase-TM-LRR	*				9	6	3:2	0.96504
AT5G49780	kinase-TM-LRR	*				11	11	1:1	1.14337
AT5G65970				*		7	24	7:24	1.84519
AT5G66610	PF00412: LIM domain		*	*	*	17	4	17:4	5.98096
AT5G66620	PF00412: LIM domain		*	*	*	21	6	7:2	1.47004
AT5G66630	NBS-LIM		*	*	*	14	7	2:1	1.45727
AT5G66640	PF00412: LIM domain		*	*	*	10	5	2;1	1.43632
AT5G67200	kinase-TM-LRR		*			2	5	2:5	0.75575

A6.2: FLOWERING TIME-LINKED GENES

Gene	Wall	Garden	Other	All UK	Ka(non-syn)	Ks (syn)	Ka/Ks ratio	Crossover rate
AT2G18790		*		*	6	50	3:25	1.37066
AT2G33835		*			4	2	2:1	0.16743
AT2G39810		*			7	6	7:6	0.61648
AT4G00650			*		10	1	10:1	0.84824
AT5G10140	*				3	1	3:1	0.43923
AT5G62640		*			3	1	3:1	0.42853